



RESEARCH CENTER

FIELD

Digital Health, Biology and Earth

Activity Report 2017

Section Scientific Foundations

Edition: 2018-02-19

COMPUTATIONAL BIOLOGY

1. ABS Project-Team	5
2. AMIBIO Team	9
3. BEAGLE Project-Team	15
4. BIGS Project-Team	20
5. BONSAI Project-Team	22
6. CAPSID Project-Team	24
7. DYLISS Project-Team	28
8. ERABLE Project-Team	35
9. GENSCALE Project-Team	40
10. IBIS Project-Team	42
11. LIFEWARE Project-Team	47
12. MORPHEME Project-Team	50
13. PLEIADE Team	52
14. SERPICO Project-Team	54
15. TAPDANCE Team	56
16. VIRTUAL PLANTS Project-Team	57

COMPUTATIONAL NEUROSCIENCE AND MEDICINE

17. ARAMIS Project-Team	59
18. ASCLEPIOS Project-Team	61
19. ATHENA Project-Team	64
20. BIOVISION Team	68
21. CAMIN Team	71
22. GALEN Project-Team	73
23. MATHNEURO Team	77
24. MIMESIS Team	79
25. MNEMOSYNE Project-Team	83
26. NEUROSYS Project-Team	87
27. PARIETAL Project-Team	89
28. VISAGES Project-Team	93

EARTH, ENVIRONMENTAL AND ENERGY SCIENCES

29. AIRSEA Project-Team	95
30. ANGE Project-Team	100
31. CASTOR Project-Team	104
32. COFFEE Project-Team	105
33. FLUMINANCE Project-Team	107
34. LEMON Team	110
35. MAGIQUE-3D Project-Team	121
36. SERENA Project-Team	127
37. STEEP Project-Team	129
38. TONUS Team	132

MODELING AND CONTROL FOR LIFE SCIENCES

39. BIOCORE Project-Team	135
40. CARMEN Project-Team	137
41. DRACULA Project-Team	140
42. M3DISIM Project-Team	143
43. MAMBA Project-Team	144
44. MONC Project-Team	155
45. MYCENAE Project-Team	162
46. NUMED Project-Team	165
47. REO Project-Team	169
48. SISTM Project-Team	172
49. XPOP Project-Team	174

ABS Project-Team

3. Research Program

3.1. Introduction

The research conducted by ABS focuses on three main directions in Computational Structural Biology (CSB), together with the associated methodological developments:

- Modeling interfaces and contacts,
- Modeling macro-molecular assemblies,
- Modeling the flexibility of macro-molecules,
- Algorithmic foundations.

3.2. Modeling interfaces and contacts

Keywords: Docking, interfaces, protein complexes, structural alphabets, scoring functions, Voronoi diagrams, arrangements of balls.

The Protein Data Bank, <http://www.rcsb.org/pdb>, contains the structural data which have been resolved experimentally. Most of the entries of the PDB feature isolated proteins⁰, the remaining ones being protein - protein or protein - drug complexes. These structures feature what Nature does – up to the bias imposed by the experimental conditions inherent to structure elucidation, and are of special interest to investigate non-covalent contacts in biological complexes. More precisely, given two proteins defining a complex, interface atoms are defined as the atoms of one protein *interacting* with atoms of the second one. Understanding the structure of interfaces is central to understand biological complexes and thus the function of biological molecules [45]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [48]. Current investigations follow two routes. From the experimental perspective [31], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [42]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [37].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change⁰, or a function stating that a complex is more stable than another one, in which case the value returned is a score and not an energy. Borrowing to statistical mechanics [26], the usual way to design scoring functions is to mimic the so-called potentials of mean force. To put it briefly, one reverts Boltzmann's law, that is, denoting $p_i(r)$ the probability of two atoms –defining type i – to be located at distance r , the (free) energy assigned to the pair is computed as $E_i(r) = -kT \log p_i(r)$. Estimating from the PDB one function $p_i(r)$ for each type of pair of atoms, the energy of a complex is computed as the sum of the energies of the pairs located within a distance threshold [46], [33]. To compare the energy thus obtained to a reference state, one may compute $E = \sum_i p_i \log p_i / q_i$, with p_i the observed frequencies, and q_i the frequencies stemming from an a priori model [38]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

Describing interfaces poses problems in two settings: static and dynamic.

⁰For structures resolved by crystallography, the PDB contains the asymmetric unit of the crystal. Determining the biological unit from the asymmetric unit is a problem in itself.

⁰The Gibbs free energy of a system is defined by $G = H - TS$, with $H = U + PV$. G is minimum at an equilibrium, and differences in G drive chemical reactions.

In the static setting, one seeks the minimalist geometric model providing a relevant bio-physical signal. A first step in doing so consists of identifying interface atoms, so as to relate the geometry and the bio-chemistry at the interface level [12]. To elaborate at the atomic level, one seeks a structural alphabet encoding the spatial structure of proteins. At the side-chain and backbone level, an example of such alphabet is that of [27]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [47], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond – a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [30].

In the dynamic setting, one wishes to understand whether selected (hot) residues exhibit specific dynamic properties, so as to serve as anchors in a binding process [41]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

3.3. Modeling macro-molecular assemblies

Keywords: Macro-molecular assembly, reconstruction by data integration, proteomics, modeling with uncertainties, curved Voronoi diagrams, topological persistence.

3.3.1. Reconstruction by Data Integration

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [25]. In a nutshell, the strategy is reminiscent from NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This strategy has been in particular used to propose plausible models of the Nuclear Pore Complex [24], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

3.3.2. Modeling with Uncertainties and Model Assessment

Reconstruction by data integration requires three ingredients. First, a parametrized model must be adopted, typically a collection of balls to model a protein with pseudo-atoms. Second, as in NMR, a functional measuring the agreement between a model and the data must be chosen. In [23], this functional is based upon *restraints*, namely penalties associated to the experimental data. Third, an optimization scheme must be selected. The design of restraints is notoriously challenging, due to the ambiguous nature and/or the noise level of the data. For example, Tandem Affinity Purification (TAP) gives access to a *pullout* i.e. a list of protein types which are known to interact with one tagged protein type, but no information on the number of complexes or on the stoichiometry of proteins types within a complex is provided. In cryo-EM, the envelope enclosing an assembly is often imprecisely defined, in particular in regions of low density. For immuno-EM labelling experiments, positional uncertainties arise from the microscope resolution.

These uncertainties coupled with the complexity of the functional being optimized, which in general is non convex, have two consequences. First, it is impossible to single out a unique reconstruction, and a set of plausible reconstructions must be considered. As an example, 1000 plausible models of the NPC were reported in [23]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

3.4. Modeling the flexibility of macro-molecules

Keywords: Folding, docking, energy landscapes, induced fit, molecular dynamics, conformers, conformer ensembles, point clouds, reconstruction, shape learning, Morse theory.

Proteins in vivo vibrate at various frequencies: high frequencies correspond to small amplitude deformations of chemical bonds, while low frequencies characterize more global deformations. This flexibility contributes to the entropy thus the free energy of the system *protein - solvent*. From the experimental standpoint, NMR studies generate ensembles of conformations, called *conformers*, and so do molecular dynamics (MD) simulations. Of particular interest while investigating flexibility is the notion of correlated motion. Intuitively, when a protein is folded, all atomic movements must be correlated, a constraint which gets alleviated when the protein unfolds since the steric constraints get relaxed⁰. Understanding correlations is of special interest to predict the folding pathway that leads a protein towards its native state. A similar discussion holds for the case of partners within a complex, for example in the third step of the *diffusion - conformer selection - induced fit* complex formation model.

Parameterizing these correlated motions, describing the corresponding energy landscapes, as well as handling collections of conformations pose challenging algorithmic problems.

At the side-chain level, the question of improving rotamer libraries is still of interest [29]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [44]. The standard methodology to analyze such a point cloud consists of resorting to normal modes. Recently, though, more elaborate methods resorting to more local analysis [40], to Morse theory [35] and to analysis of meta-stable states of time series [36] have been proposed.

3.5. Algorithmic foundations

Keywords: Computational geometry, computational topology, optimization, data analysis.

Making a stride towards a better understanding of the biophysical questions discussed in the previous sections requires various methodological developments, which we briefly discuss now.

3.5.1. Modeling Interfaces and Contacts

In modeling interfaces and contacts, one may favor geometric or topological information.

On the geometric side, the problem of modeling contacts at the atomic level is tantamount to encoding multi-body relations between an atom and its neighbors. On the one hand, one may use an encoding of neighborhoods based on geometric constructions such as Voronoi diagrams (affine or curved) or arrangements of balls. On the other hand, one may resort to clustering strategies in higher dimensional spaces, as the p neighbors of a given atom are represented by $3p - 6$ degrees of freedom – the neighborhood being invariant upon rigid motions. The information gathered while modeling contacts can further be integrated into interface models.

On the topological side, one may favor constructions which remain stable if each atom in a structure *retains* the same neighbors, even though the 3D positions of these neighbors change to some extent. This process is observed in flexible docking cases, and call for the development of methods to encode and compare shapes undergoing tame geometric deformations.

3.5.2. Modeling Macro-molecular Assemblies

In dealing with large assemblies, a number of methodological developments are called for.

⁰ Assuming local forces are prominent, which in turn subsumes electrostatic interactions are not prominent.

On the experimental side, of particular interest is the disambiguation of proteomics signals. For example, TAP and mass spectrometry data call for the development of combinatorial algorithms aiming at unraveling pairwise contacts between proteins within an assembly. Likewise, density maps coming from electron microscopy, which are often of intermediate resolution (5-10Å) call the development of noise resilient segmentation and interpretation algorithms. The results produced by such algorithms can further be used to guide the docking of high resolutions crystal structures into maps.

As for modeling, two classes of developments are particularly stimulating. The first one is concerned with the design of algorithms performing reconstruction by data integration, a process reminiscent from non convex optimization. The second one encompasses assessment methods, in order to single out the reconstructions which best comply with the experimental data. For that endeavor, the development of geometric and topological models accommodating uncertainties is particularly important.

3.5.3. Modeling the Flexibility of Macro-molecules

Given a sampling on an energy landscape, a number of fundamental issues actually arise: how does the point cloud describe the topography of the energy landscape (a question reminiscent from Morse theory)? Can one infer the effective number of degrees of freedom of the system over the simulation, and is this number varying? Answers to these questions would be of major interest to refine our understanding of folding and docking, with applications to the prediction of structural properties. It should be noted in passing that such questions are probably related to modeling phase transitions in statistical physics where geometric and topological methods are being used [39].

From an algorithmic standpoint, such questions are reminiscent of *shape learning*. Given a collection of samples on an (unknown) *model*, *learning* consists of guessing the model from the samples – the result of this process may be called the *reconstruction*. In doing so, two types of guarantees are sought: topologically speaking, the reconstruction and the model should (ideally!) be isotopic; geometrically speaking, their Hausdorff distance should be small. Motivated by applications in Computer Aided Geometric Design, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison; the study of Morse-like constructions stemming from distance functions to points; the analysis of topological invariants of the model and the samples, and their comparison.

AMIBIO Team

3. Research Program

3.1. RNA and protein structures

At the secondary structure level, we contributed novel generic techniques applicable to dynamic programming and statistical sampling, and applied them to design novel efficient algorithms for probing the conformational space. Another originality of our approach is that we cover a wide range of scales for RNA structure representation. For each scale (atomic, sequence, secondary and tertiary structure...) cutting-edge algorithmic strategies and accurate and efficient tools have been developed or are under development. This offers a new view on the complexity of RNA structure and function that will certainly provide valuable insights for biological studies.

3.1.1. Discrete representations and complexity

Participants: Yann Ponty, Wei Wang, Antoine Soulé, Juraj Michalik.

Common activity with J. Waldispühl (McGill) and A. Denise (LRI).

Ever since the seminal work of Zuker and Stiegler, the field of RNA bioinformatics has been characterized by a strong emphasis on the secondary structure. This discrete abstraction of the 3D conformation of RNA has paved the way for a development of quantitative approaches in RNA computational biology, revealing unexpected connections between combinatorics and molecular biology. Using our strong background in enumerative combinatorics, we propose generic and efficient algorithms, both for sampling and counting structures using dynamic programming. These general techniques have been applied to study the sequence-structure relationship [46], the correction of pyrosequencing errors [38], and the efficient detection of multi-stable RNAs (riboswitches) [42], [43].

Increasingly, we develop and study parameterized complexity approaches, based on dynamic programming over a tree decomposition, for several combinatorial problems, including RNA design, structure-sequence alignment (aka threading in the context of proteins). The later problem is at the core of Wei Wang's Phd, successfully defended in Dec 2017. In the context of our probabilistic approaches, often based on random generation, such parameterized algorithms usually follow proofs of hardness for the associated enumeration problems.

3.1.2. RNA design.

Participants: Alice Héliou, Yann Ponty.

Joint project with A. Denise (sc Lri), J. Waldispühl (McGill), D. Barash (Univ. Ben-Gurion), and C. Chauve (Simon Fraser University).

It is a natural pursuit to build on our understanding of the secondary structure to construct artificial RNAs performing predetermined functions, ultimately targeting therapeutic and synthetic biology applications. Towards this goal, a key element is the design of RNA sequences that fold into a predetermined secondary structure, according to established energy models (inverse-folding problem). Quite surprisingly, and despite two decades of studies of the problem, the computational complexity of the inverse-folding problem is currently unknown.

Within AMIBio, we develop a new methodology, based on weighted random generation [25] and multidimensional Boltzmann sampling, for this problem. Initially lifting the constraint of folding back into the target structure, we explored the random generation of sequences that are compatible with the target, using a probability distribution which favors exponentially sequences of high affinity towards the target. A simple posterior rejection step selects sequences that effectively fold back into the latter, resulting in a *global sampling* pipeline that showed comparable performances to its competitors based on local search [32].

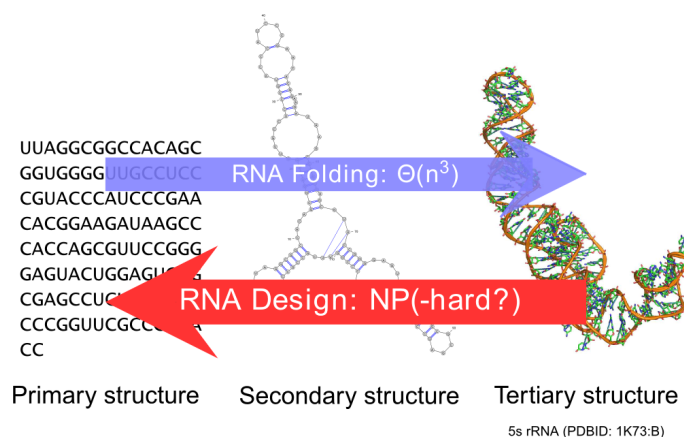


Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

The main advantages of this approach is its linear complexity, and its flexibility in incorporating constraints. Indeed, extensive experiments revealed a drift of existing software towards sequences of high G+C-content, and we showed how to control this distributional bias by using multidimensional Boltzmann sampling [37], [36]. Recently, we are extending this approach to the design of RNAs with multiple structures, developing a Fixed-Parameter Tractable framework that naturally extends to capture negative design goals.

3.1.3. Modeling large macromolecular architectures

Participants: Yann Ponty, Afaf Saaidi, Mireille Régnier, Amélie Héliou.

Joint projects with A. Denise (LRI), D. Barth (Versailles), J. Cohen (Paris-Sud), B. Sargueil (Paris V) and Jérôme Waldispühl (McGill).

The modeling of large RNA 3D structures, that is predicting the three-dimensional structure of a given RNA sequence, relies on two complementary approaches. The approach by homology is used when the structure of a sequence homologous to the sequence of interest has already been resolved experimentally. The main problem then is to calculate an alignment between the known structure and the sequence. The ab initio approach is required when no homologous structure is known for the sequence of interest (or for some parts of it). We contribute methods inspired by both of these directions.

We also develop homology-based approaches for structure modeling, and developed a general setting for the problem of RNA structure-sequence alignment, known to be NP-hard in the presence of complex topological features named pseudoknots (PKs). Our approach is based on tree decomposition of structures and gives rises to a general parameterized algorithm, where the exponential part of the complexity depends on the family of structures [39]. This work unifies and generalizes a number of recent works on specific families, and enables the curation of multiple alignments for RNA families featuring PKs, correcting certain bias introduced by PK-oblivious methods.

3.2. Séquences

Participants: Mireille Régnier, Philippe Chassignet, Yann Ponty, Jean-Marc Steyaert, Alice Héliou, Antoine Soulé.

String searching and pattern matching is a classical area in computer science, enhanced by potential applications to genomic sequences. In CPM/SPIRE community, a focus is given to general string algorithms and associated data structures with their theoretical complexity. Our group specialized in a formalization based on languages, weighted by a probabilistic model. Team members have a common expertise in enumeration and random generation of combinatorial sequences or structures, that are *admissible* according to some given constraints. A special attention is paid to the actual computability of formula or the efficiency of structures design, possibly to be reused in external software.

As a whole, motif detection in genomic sequences is a hot subject in computational biology that allows to address some key questions such as chromosome dynamics or annotation. Among specific motifs involved in molecular interactions, one may cite protein-DNA (cis-regulation), protein-protein (docking), RNA-RNA (miRNA, frameshift, circularisation). This area is being renewed by high throughput data and assembly issues. New constraints, such as energy conditions, or sequencing errors and amplification bias that are technology dependent, must be introduced in the models. A collaboration has been established with LOB, at Ecole Polytechnique, who bought a sequencing machine, through the co-advised thesis of Alice Héliou. An other aim is to combine statistical sampling with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [34]. In general, in the future, our methods for sampling and sequence data analysis should be extended to take into account such constraints, that are continuously evolving.

3.2.1. Combinatorial Algorithms and motifs

Participants: Mireille Régner, Philippe Chassignet, Alice Héliou.

Besides applications [41] of analytic combinatorics to computational biology problems, the team addressed general combinatorial problems on words and fundamental issues on languages and data structures. Motif detection combines an algorithmic search of potential sites and a significance assessment. To assess the significance of an observation usually requires the evaluation of a quantitative criterion such as the P-value. In the recent years, a general scheme of derivation of analytic formula for the P-value under different constraints (k -occurrence, first occurrence, overrepresentation in large sequences,...) has been provided. It relies on a representation of continuous sequences of overlapping words, currently named *clumps* or *clusters* in a graph [35]. Recursive equations to compute p -values may be reduced to a traversal of that graph, leading to a linear algorithm. This improves over the space and time complexity of the generating function approach or previous probabilistic weighted automata.

In [45], it is claimed that half of the genome consists of different types of repeats. One may cite microsatellites, DNA transposons, transposons, long terminal repeats (LTR), long interspersed nuclear elements (LINE), ribosomal DNA, short interspersed nuclear elements (SINE). Therefore, knowledge about the length of repeats is a key issue in several genomic problems, notably assembly or re-sequencing. Preliminary theoretical results are given in [29], and, recently, heuristics have been proposed and implemented [26], [40], [23]. A dual problem is the length of minimal absent words. Minimal absent words are words that do not occur but whose proper factors all occur in the sequence. Their computation is extremely related to finding maximal repeats (repeat that can not be extended on the right nor on the left). The comparison of the sets of minimal absent words provides a fast alternative for measuring approximation in sequence comparison [22], [24].

Recently, it was shown that considering the words which occur in one sequence but do not in another can be used to detect biologically significant events [44]. We have studied the computation of minimal absent words and we have provided new linear implementations [20]. We are now working on a dynamic approach to compute minimal absent words for a sliding window. For a sequence of size n , we expect a complexity of $O(n)$ time and space, independent of the size of the window. This approach could be used to align a sequence on a larger sequence using minimal absent words for comparison.

3.2.2. Random generation

Participants: Yann Ponty, Juraj Michalik, Christelle Rovetta.

Analytical methods may fail when both sequential and structural constraints of sequences are to be modelled or, more generally, when molecular *structures* such as RNA structures have to be handled. The random generation of combinatorial objects is a natural, alternative, framework to assess the significance of observed phenomena. General and efficient techniques have been developed over the last decades to draw objects uniformly at random from an abstract specification. However, in the context of biological sequences and structures, the uniformity assumption becomes unrealistic, and one has to consider non-uniform distributions in order to derive relevant estimates. Typically, context-free grammars can handle certain kinds of long-range interactions such as base pairings in secondary RNA structures.

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [27]: instead of formulating the problem as a classic optimization, this new approach uses statistical sampling within the space of solutions. Besides giving better, more robust, results, it allows for a fruitful adaptation of tools and algorithms derived in a purely combinatorial setting.

We also introduced algorithms and data structures for a non-redundant generation of combinatorial objects. In situations where the search space of a problem can be unambiguously explored using dynamic programming, such algorithms generate objects within a postulated distribution, conditioned to avoid previously generated objects. This method can be used to probe objects having lower probabilities, a desirable property in the context of RNA kinetics studies, or could lead to better estimators in context where the exact emission probability of each object can be computed.

3.3. 3D interaction and structure prediction

Participant: Amélie Héliou.

The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. This is specially challenging as structure flexibility is key and multi-scale modelling [21], [28] and efficient code are essential [33].

Our project covers various aspects of biological macromolecule structure and interaction modelling and analysis. First protein structure prediction is addressed through combinatorics. The dynamics of these types of structures is also studied using statistical and robotics inspired strategies. Both provide a good starting point to perform 3D interaction modelling, accurate structure and dynamics being essential.

Our group benefits from a good collaboration network, mainly at Stanford University (USA), HKUST (Hong-Kong) and McGill (Canada). The computational expertise in this field of computational structural biology is represented in a few large groups in the world (e.g. Pande lab at Stanford, Baker lab at U.Washington) that have both dry and wet labs. At Inria, our interest for structural biology is shared by the ABS and ORPAILLEUR project-teams. Our activities are however now more centered around protein-nucleic acid interactions, multi-scale analysis, robotics inspired strategies and machine learning than protein-protein interactions, algorithms and geometry. We also shared a common interest for large biomolecules and their dynamics with the NANO-D project team and their adaptative sampling strategy. As a whole, we contribute to the development of geometric and machine learning strategies for macromolecular docking.

Game theory was used by M. Boudard in her PhD thesis, defended in 2015, to predict the 3d structure of RNA. In her PhD thesis, co-advised by J. Cohen (LRI), A. Héliou extended the approach to the prediction of protein structures.

3.3.1. Robotics-inspired structure and dynamics

Participant: Amélie Héliou.

We recently work on a robotics approach to sample the conformational space of macromolecules like RNAs [1]. The robotics approach allows maintaining the secondary structure of the RNA fixed, as an unfolding is very unlikely and energetically demanding. By this approach we also dramatically reduce the number of degrees of freedom in the molecule. The conformational space becomes possible to be sampled. This reduction does not reduce the quality of the sampling.

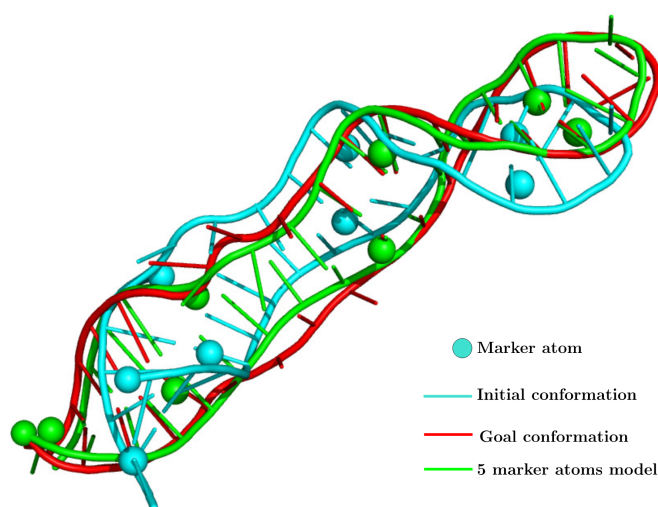


Figure 2. The cyan structure is the initial conformation, the red structure is the goal conformation. The full-atom initial conformation was driven toward the goal conformation using only the position of the goal sphere atoms. The green conformation is the result obtained; spheres perfectly overlap with the goal position and the overall conformation is really close to the goal conformation.

Our current work consists in applying the same approach to a targeted move. The motion is then driven either by the position of a few atoms or the distances between couple of atoms. These two aspects are under development and will increase our capacity to integrate experimental data. Our method can drive a RNA conformation toward another conformation of the same RNA given only the position of a few atoms (marker atoms).

For instance double electron-electron resonance (DEER) experimental results are distributions of distances. Probes are attached to the molecules and the distances between to probes is measured and outputted as a distribution. Our method is able to sample an ensemble of all-atom conformations that can explain the distance distribution.

3.3.2. Game theory and molecular folding

Participant: Amélie Héliou.

Building on a previous collaboration with LRI (J. Cohen, expert in algorithmic game theory – GALAC team), we have extended an original approach to structure prediction based on game theory, previously introduced as a proof-of-concept [31]. Our model of the folding model as a game represents an alternative paradigm to thermodynamics and energy minimization, in which individual residues are considered as selfish players, whose strategy (position/orientation) is driven by a desire to increase their local utility function (consistency with observed local conformations). The conformations where residues can no longer gain any utility by changing their local conformations are called Nash equilibria. They can be thought of as local minima of a suitably-defined energy function, and can sometimes be efficiently computed using stochastic strategies.

Our work is first to find an algorithm that can guarantee the convergence to a Nash equilibrium (a state where no player would increase his payoff by playing something different alone) and prove their convergence. At the same time, we are looking for efficient and biologically relevant ways of defining the game settings so that Nash equilibria correspond to folded states. One direction would be to draw a parallel between Nash equilibria

and local minima of the kinetic landscape. This line of research also raises questions related to learning Nash equilibria.

BEAGLE Project-Team

3. Research Program

3.1. Introduction

As stated above, the research topics of the BEAGLE Team are centered on the modelization and simulation of cellular processes. More specifically, we focus on two specific processes that govern cell dynamics and behavior: Evolution and Biophysics.

3.2. Research axis 1: Computational cellular biochemistry

3.2.1. *Biochemical kinetics with unconventional diffusion.*

The movement of many biomolecules such as proteins in living cells has been reported as spatially heterogeneous diffusion (position-dependent diffusion coefficient) or even as anomalous diffusion, whereby the mean-squared displacement scales sub-linearly with time, $\langle \mathbf{r}^2 \rangle \sim t^\alpha$ with $\alpha < 1$. The influences of such deviations from simple Brownian motion on the biochemical reactions that take place in these media are just starting to be explored. Part of our efforts was aimed at improving the modeling of sub-diffusion in an intracellular context by extending the existing models of obstacle-based sub-diffusion to mobile obstacles [37] or proposing age-structured PDEs to model sub-diffusion [38], [43]. We also explored the effects of locally slowed-down diffusion or local sub-diffusion on the dynamics of simple reactions such as the ligand-binding equilibrium or protein aggregation [44], [39], [77], [78], [47].

3.2.2. *Spatial and temporal organization of gene expression.*

Gene expression is highly organized in space and diffusing factors often have to reach and bind partners which positions are not well-mixed, but exhibit stationary spatial organization. We have studied how the spatial organization of genes can affect their interaction [64], [65]. Combining modeling and experiments, we have studied the stochasticity of gene expression in single metazoan cells (chicken cells) and provided its quantification via stochasticity measures [89], [33], [48].

3.2.3. *Adipose tissue modeling*

Most cellular pathway models and equations assume that the cell / medium is size / volume constant during the whole duration of the process. However there is an important variability in size and volume within tissue, within cell type and of course during the life of a single cell. We have studied the mechanism of lipid storage in adipocytes both at equilibrium and non-equilibrium [80], [79] and confirmed that the main model predictions are indeed present in rats [57], [58].

3.2.4. *Current Objectives*

We will intensify our collaborations with mathematicians and theoretical physicists in order to develop age-structured PDEs as models of subdiffusion and (sub)diffusion-reaction coupling in the perspectives of non-parametric inference for classification of intracellular trajectories or the modeling of subdiffusion-reaction coupling. Regarding experimental data, we are developing collaborations with experts of super-resolution microscopy in order to be able to combine our models with single-molecule trajectory data and spatial localization data at single-molecule resolution in living cells. Moreover, since September 2017, Carole Knibbe shares her research time (co-affiliation) between Beagle and the biomedical research laboratory CarMeN (<http://carmen.univ-lyon1.fr>). On this occasion, her research has joined Research axis 1. In collaboration with experimental biologists of the CarMeN laboratory (e.g. Marie-Caroline Michalski), she now works on mathematical and computational models of lipolysis kinetics, with the aim of reaching a quantitative understanding of how the spatial supramolecular organization of lipids influences their digestion. A variety of experimental data are available at CarMeN to calibrate the models at the various scales, from in vitro enzymatic data, to cultures of intestinal cells, to animal models, to clinical data in normal and obese patients.

3.3. Research axis 2: Models for Molecular Evolution

3.3.1. *Backwards models. The example of dating with transfers.*

Here the goal is to compare extant genomes in order to propose plausible scenarios for their evolution. We constructed a multi-scale model of evolution, at the levels of species, chromosomes, genes and nucleotides. Reconstruction algorithms proposed a catalog of putative ancient transfers in different domains of life, as cyanobacteria, fungi, archaea. As transfers can only occur between contemporaneous species, they contain an information on the timing of life diversification. This allowed us to exploit a completely novel and abundant information for the history of biodiversity [81], [83], [82], [84], [45].

3.3.2. *Forward models I. In silico experimental evolution.*

We have made several crucial advances in particular around the *Aevol* software that we have been developing for more than 10 years. This software consists in evolving *in silico* a population of digital genomes from an ancestral state and quantifying expected or unexpected systematic behaviors, in function of settled parameters (see Figure 1). We formalized the method [35] and implemented a tracking system that allows us to measure histories of genes [61]. We explored the potentiality of epigenetic inheritance through R-Aevol, an extension of Aevol implementing genetic regulation and epigenetic (cytoplasmic) inheritance [86], [87]. We also developed models with a new artificial chemistry in the context of the European project Evoevo (see below).

3.3.3. *Forward models II. Theoretical investigations.*

Simulations are made when models are too complicated to be handled analytically. However, it is sometimes worth simplifying the model in order to demonstrate a given property and have a proven behavior in function of perfectly identified parameters. We have obtained important results in this scope, mostly about the size of impact of selection on genome sizes [55]. Another set of theoretical results was obtained by the introduction of phenotypic noise in Fisher's Geometric Model (see below).

3.3.4. *The Forward-Backward and the double blind procedure.*

One objective of simulations or mathematical forward models is to explain the shape of some extant particular data. When one (or more) evolutionary mechanism is evoked to explain a given dataset, simulations can test if the invoked mechanism is indeed likely to produce the observed dataset. We investigated for example the causes of genome reduction in *Prochlorococcus* and compared them with genome reduction in endosymbionts [35], [36]. Another way we made the forward and backward models enrich each other is by using forward models to construct test instances for the backward methods. We have developed forward and backward models that are not scientifically linked, except by a common biological background [42], [41]. That led us to initiate the development of a new version of the Aevol software, "Aevol_ACGT", that will include a four nucleotides genome decoded through a realistic genetic code [46], [63].

3.3.5. *Current Objectives*

We are improving and publicizing the dating method using gene transfers. We are leaders on this subject, published already in very good journals or labels, so we still see this subject as highly promising. A second important goal is producing benchmarks for the validation of evolutionary studies. This can be very useful for the evolutionary biology community, which lacks good validation standards. We provided some proofs of concept during the last years and will work towards a completely usable platform in the following years. Moreover we are pushing our *in silico* experimental model towards standard models of theoretical biology to test the robustness of some hypotheses to complex models. Finally we want to explore the possibility to co-construct biological knowledge with the general public through participative platforms. We will propose a protocol usable by the public in order to solve unknown parts of the tree of life, which size now calls for a participative project like encyclopedias or maps. This will demand new skills in computer science, participative sciences, and also will depend on consequent funding applications.

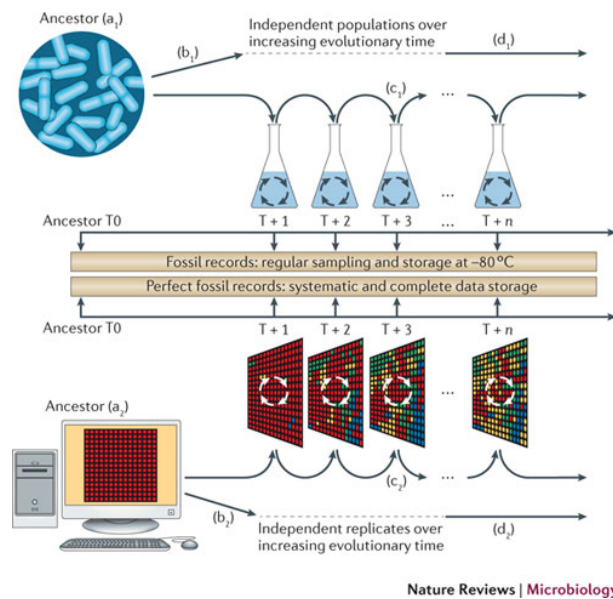


Figure 1. Parallel between experimental evolution and artificial evolution

3.4. Research axis 3: Computational systems biology of neurons and astrocytes

3.4.1. Understanding neuronal plasticity

One of our main achievements concerns the study of synaptic plasticity in the the basal ganglia, the locus of procedural learning (the learning of skills and habits) and reinforcement learning. We have been carrying out joint experimental-modeling research to decipher the signaling pathways that underlay synaptic plasticity in this brain structure [49], [50]. This result is a significant advance for the understanding of synaptic plasticity because it shows that plasticity can be triggered even with low activity levels. In related threads of research, we have also studied the molecular basis of intrinsic plasticity [66] or the adult neurogenesis [66].

3.4.2. Biophysical models of astrocyte signaling

Neurons represent roughly half of the brain cells. The other part is made of glial cells, that include *e.g.* microglia, oligodendrocytes or astrocytes. The role of glial cells in cognition, learning or memory is a new field of research in neuroscience. We have developed reference mathematical models in the field, focusing on intra- and inter-cellular astrocyte calcium signals and astrocyte-to-neurons signaling via neuro- and gliotransmitters [67], [52], [62], [90]. We have initiated a modeling and theoretical study of the modulation of neuronal plasticity by astrocyte signaling [51].

3.4.3. Motor model for perceptive neural coding

In animal acoustic perception, the canonical representation of sound is through Fourier decomposition via a spectrographic representation. While generic, this approach does not take into account the (probable) tuning of neurons to familiar sound and therefore ad-hoc sound statistics. We have proposed an alternative view where sound representation is based on the motor representation needed to produce the acoustic output. In the context of songbirds, we proposed mechanical models that reproduce bird vocalization using muscles contractions dynamics. Part of auditory neural tuning is probably focused on decoding vocalization of endogenous individuals in the context of a social network. We have developed a series of analysis tools to

identify acoustic social networks of Zebra Finches and applied them to the study of small bird groups in the lab [70], [71], [88], [53], [54].

3.4.4. Current Objectives

Research axis 3 will continue but with considerable modifications. We are carrying on work on synaptic plasticity in the basal ganglia focusing on the modulation of synaptic plasticity by neuromodulators like dopamine. In particular, understanding the impact of dopamine signaling in the basal ganglia could provide us with significant advances in dopamine-related pathologies (Parkinson's, addictions...). Regarding the work on astrocytes, we will concentrate our efforts on the deciphering of calcium signaling in the finest astrocyte branchlets that are expected to be the locus of interactions with the neurons. Our objective is to uncover how the spatial organization of the implied signaling molecules impacts the signaling between the synapse and the astrocyte. To obtain experimental data related to this spatial intracellular organization, we will initiate several new collaborations with experimental neuroscientists.

3.5. Research axis 4: EvoEvo

The EvoEvo project was a highly interdisciplinary project ranging from wet experiments (experimental evolution) to software application (evolving a musical personal companion).

3.5.1. Computational experiments

To help understand the results of the wet experiments conducted by our partners, we designed several computational experiments that used the models developed by the team (see Research axis 2 above).

Dynamics of innovation in viral strains. Viral strains show a surprising evolutionary dynamics in which long periods of stasis are interrupted by short evolutionary bursts. To understand this dynamics, we launched a very large scale *in silico* experiment in the Aevol software [40], [60].

Dynamics of mutator strains. The mutation rate of bacterial strains is known to adapt to the evolutionary conditions through the emergence of “mutator strains” that have a 10 to 100 fold raised mutation rate compared to the wild-type. Mutator strains are supposed to raise in specific conditions such as stress or environmental change. We studied this situation in the Aevol software and showed the importance of the reorganization of the genome structure in this phenomenon [76].

Diversification in seasonal environments. One of the targets of the EvoEvo project was the evolution of open-endedness. As a first step, we provided a careful definition of open-endedness [34], [85] and developed a new model, Evo2Sim (<http://www.evoevo.eu/evo2sim/>) [72], that contains all the ingredients to enable the emergence of structured populations. We then used it to simulate the evolution of populations in batch cultures and continuous conditions [73], [74], [75].

3.5.2. Bioinspired computation

Once EvoEvo strategies were identified *in vivo* and *in silico*, the final objective of the EvoEvo project was to transfer them into the ICT domain by developing new evolutionary metaheuristics and applying them to real problems.

Evolutionary-based subspace clustering. One of the main outcomes of the *in silico* models was the importance of chromosomal rearrangements (and of their consequences on the genome organization) for the evolutionary process. Hence, we proposed to use these properties to develop new evolutionary algorithms specifically dedicated to the subspace clustering task [69], [68]. We also developed the first evolutionary algorithm dedicated to the subspace clustering of datastreams.

EvoMove: A musical personal companion. To validate our evolutionary-based subspace clustering algorithm, we used it to develop EvoMove, a musical personal companion to be used by dancers and musicians [12], [24]. EvoMove has been tested by professional dancers (<http://www.evoevo.eu/evomove-working-session/>) and has been presented to the public in the context of the “Meute” (“Herd”) performance that has been played eight times between February and July 2017.

3.5.3. Current Objectives

Research axis 4 is relatively new in the team since it started with the EvoEvo European project. Hence many of the activities of this research axis are still fully active. We will carry on the cooperation on the innovative dynamics of viral strains, with the objective to study whether the dynamics we observed in viruses is also observed in other kinds of genomes. We also will investigate further the role of large scale events (chromosomal rearrangements, recombination...) on the macro-evolutionary dynamics and on the interaction between micro- and macro-evolution. Finally, we will carry on our work on evolutionary-inspired subspace clustering algorithms and on the evolutionary metaphor applied to the development of new ICT technologies with a specific focus on the integration of temporal data.

BIGS Project-Team

3. Research Program

3.1. Introduction

We give here the main lines of our research that belongs to the domains of probability and statistics. For a better understanding, we made the choice to structure them in four items. Although this choice was not arbitrary, the outlines between these items are sometimes fuzzy because each of them deals with modeling and inference and they are all interconnected.

3.2. Stochastic modeling

Our aim is to propose relevant stochastic frameworks for the modeling and the understanding of biological systems. The stochastic processes are particularly suitable for this purpose. Among them, Markov chains give a first framework for the modeling of population of cells [79], [56]. Piecewise deterministic processes are non diffusion processes also frequently used in the biological context [45], [55], [47]. Among Markov model, we also developed strong expertise about processes derived from Brownian motion and Stochastic Differential Equations [71], [54]. For instance, knowledge about Brownian or random walk excursions [78], [70] helps to analyse genetic sequences and to develop inference about it. However, nature provides us with many examples of systems such that the observed signal has a given Hölder regularity, which does not correspond to the one we might expect from a system driven by ordinary Brownian motion. This situation is commonly handled by noisy equations driven by Gaussian processes such as fractional Brownian motion or fractional fields. The basic aspects of these differential equations are now well understood, mainly thanks to the so-called *rough paths* tools [62], but also invoking the Russo-Vallois integration techniques [72]. The specific issue of Volterra equations driven by fractional Brownian motion, which is central for the subdiffusion within proteins problem, is addressed in [46]. Many generalizations (Gaussian or not) of this model have been recently proposed for some Gaussian locally self-similar fields, or for some non-Gaussian models [59], or for anisotropic models [42].

3.3. Estimation and control for stochastic processes

We develop inference about stochastic processes that we use for modeling. Control of stochastic processes is also a way to optimise administration (dose, frequency) of therapy.

There are many estimation techniques for diffusion processes or coefficients of fractional or multifractional Brownian motion according to a set of observations [58], [37], [44]. But, the inference problem for diffusions driven by a fractional Brownian motion has been in its infancy. Our team has a good expertise about inference of the jump rate and the kernel of Piecewise Deterministic Markov Processes (PDMP) [34], [35], [33], [36]. However, there are many directions to go further into. For instance, previous works made the assumption of a complete observation of jumps and mode, that is unrealistic in practice. We tackle the problem of inference of "Hidden PDMP". About pharmacokinetics modeling inference, we want to take into account for presence of timing noise and identification from longitudinal data. We have expertise on this subjects [38], and we also used mixed models to estimate tumor growth [39].

We consider the control of stochastic processes within the framework of Markov Decision Processes [69] and their generalization known as multi-player stochastic games, with a particular focus on infinite-horizon problems. In this context, we are interested in the complexity analysis of standard algorithms, as well as the proposition and analysis of numerical approximate schemes for large problems in the spirit of [41]. Regarding complexity, a central topic of research is the analysis of the Policy Iteration algorithm, which has made significant progress in the last years [81], [68], [53], [77], but is still not fully understood. For large problems, we have a long experience of sensitivity analysis of approximate dynamic programming algorithms for Markov Decision Processes [75], [74], [76], [61], [73], and we currently investigate whether/how similar ideas may be adapted to multi-player stochastic games.

3.4. Algorithms and estimation for graph data

A graph data structure consists of a set of nodes, together with a set of pairs of these nodes called edges. This type of data is frequently used in biology because they provide a mathematical representation of many concepts such as biological structures and networks of relationships in a population. Some attention has recently been focused in the group on modeling and inference for graph data.

Network inference is the process of making inference about the link between two variables taking into account the information about other variables. [80] gives a very good introduction and many references about network inference and mining. Many methods are available to infer and test edges in Gaussian Graphical models [80], [63], [50], [51]. However, when dealing with abundance data, because inflated zero data, we are far from gaussian assumption and we want to develop inference in this case.

Among graphs, trees play a special role because they offer a good model for many biological concepts, from RNA to phylogenetic trees through plant structures. Our research deals with several aspects of tree data. In particular, we work on statistical inference for this type of data under a given stochastic model. We also work on lossy compression of trees via linear directed acyclic graphs. These methods enable us to compute distances between tree data faster than from the original structures and with a high accuracy.

3.5. Regression and machine learning

Regression models or machine learning aim at inferring statistical links between a variable of interest and covariates. In biological study, it is always important to develop adapted learning methods both in the context of *standard* data and also for data of high dimension (with sometimes few observations) and very massive or online data.

Many methods are available to estimate conditional quantiles and test dependencies [67], [57]. Among them we have developed nonparametric estimation by local analysis via kernel methods [48], [49] and we want to study properties of this estimator in order to derive a measure of risk like confidence band and test. We study also many other regression models like survival analysis, spatio temporal models with covariates. Among the multiple regression models, we want to develop omnibus test that examine several assumptions together.

Concerning the analysis of high dimensional data, our view on the topic relies on the *French data analysis school*, specifically on Factorial Analysis tools. In this context, stochastic approximation is an essential tool [60], which allows one to approximate eigenvectors in a stepwise manner [66], [64], [65]. BIGS aims at performing accurate classification or clustering by taking advantage of the possibility of updating the information "online" using stochastic approximation algorithms [43]. We focus on several incremental procedures for regression and data analysis like linear and logistic regressions and PCA.

We also focus on the biological context of high-throughput bioassays in which several hundreds or thousands of biological signals are measured for a posterior analysis. We have to account for the inter-individual variability within the modeling procedure. We aim at developing a new solution based on an ARX (Auto Regressive model with eXternal inputs) model structure using the EM (Expectation-Maximisation) algorithm for the estimation of the model parameters.

BONSAI Project-Team

3. Research Program

3.1. Sequence processing for Next Generation Sequencing

As said in the introduction of this document, biological sequence analysis is a foundation subject for the team. In the last years, sequencing techniques have experienced remarkable advances with Next Generation Sequencing (NGS), that allow for fast and low-cost acquisition of huge amounts of sequence data, and outperforms conventional sequencing methods. These technologies can apply to genomics, with DNA sequencing, as well as to transcriptomics, with RNA sequencing. They promise to address a broad range of applications including: Comparative genomics, individual genomics, high-throughput SNP detection, identifying small RNAs, identifying mutant genes in disease pathways, profiling transcriptomes for organisms where little information is available, researching lowly expressed genes, studying the biodiversity in metagenomics. From a computational point of view, NGS gives rise to new problems and gives new insight on old problems by revisiting them: Accurate and efficient remapping, pre-assembling, fast and accurate search of non exact but quality labeled reads, functional annotation of reads, ...

3.2. Noncoding RNA

Our expertise in sequence analysis also applies to noncoding RNA. Noncoding RNA plays a key role in many cellular processes. First examples were given by microRNAs (miRNAs) that were initially found to regulate development in *C. elegans*, or small nucleolar RNAs (snoRNAs) that guide chemical modifications of other RNAs in mammals. Hundreds of miRNAs are estimated to be present in the human genome, and computational analysis suggests that more than 20% of human genes are regulated by miRNAs. To go further in this direction, the 2007 ENCODE Pilot Project provides convincing evidence that the Human genome is pervasively transcribed, and that a large part of this transcriptional output does not appear to encode proteins. All those observations open a universe of “RNA dark matter” that must be explored. From a combinatorial point of view, noncoding RNAs are complex objects. They are single stranded nucleic acid sequences that can fold forming long-range base pairings. This implies that RNA structures are usually modeled by complex combinatorial objects, such as ordered labeled trees, graphs or arc-annotated sequences.

3.3. Genome structures

Our third application domain is concerned with the structural organization of genomes. Genome rearrangements are able to change genome architecture by modifying the order of genes or genomic fragments. The first studies were based on linkage maps and fifteen year old mathematical models. But the usage of computational tools was still limited due to the lack of data. The increasing availability of complete and partial genomes now offers an unprecedented opportunity to analyze genome rearrangements in a systematic way and gives rise to a wide spectrum of problems: Taking into account several kinds of evolutionary events, looking for evolutionary paths conserving common structure of genomes, dealing with duplicated content, being able to analyze large sets of genomes even at the intraspecific level, computing ancestral genomes and paths transforming these genomes into several descendant genomes.

3.4. Nonribosomal peptides

Lastly, the team has been developing for several years a tight collaboration with ProBioGEM team in Institut Charles Viollette on nonribosomal peptides, and has become a leader on that topic. Nonribosomal peptide synthesis produces small peptides not going through the central dogma. As the name suggests, this synthesis uses neither messenger RNA nor ribosome but huge enzymatic complexes called Nonribosomal peptide synthetases (NRPSs). This alternative pathway is found typically in bacteria and fungi. It has been described

for the first time in the 70's. For the last decade, the interest in nonribosomal peptides and their synthetases has considerably increased, as witnessed by the growing number of publications in this field. These peptides are or can be used in many biotechnological and pharmaceutical applications (e.g. anti-tumors, antibiotics, immuno-modulators).

CAPSID Project-Team

3. Research Program

3.1. Classifying and Mining Protein Structures and Protein Interactions

3.1.1. Context

The scientific discovery process is very often based on cycles of measurement, classification, and generalisation. It is easy to argue that this is especially true in the biological sciences. The proteins that exist today represent the molecular product of some three billion years of evolution. Therefore, comparing protein sequences and structures is important for understanding their functional and evolutionary relationships [66], [41]. There is now overwhelming evidence that all living organisms and many biological processes share a common ancestry in the tree of life. Historically, much of bioinformatics research has focused on developing mathematical and statistical algorithms to process, analyse, annotate, and compare protein and DNA sequences because such sequences represent the primary form of information in biological systems. However, there is growing evidence that structure-based methods can help to predict networks of protein-protein interactions (PPIs) with greater accuracy than those which do not use structural evidence [45], [71]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [30].

3.1.2. Quantifying Structural Similarity

Often, proteins may be divided into modular sub-units called domains, which can be associated with specific biological functions. Thus, a protein domain may be considered as the evolutionary unit of biological structure and function [70]. However, while it is well known that the 3D structures of protein domains are often more evolutionarily conserved than their one-dimensional (1D) amino acid sequences, comparing 3D structures is much more difficult than comparing 1D sequences. However, until recently, most evolutionary studies of proteins have compared and clustered 1D amino acid and nucleotide sequences rather than 3D molecular structures.

A pre-requisite for the accurate comparison of protein structures is to have a reliable method for quantifying the structural similarity between pairs of proteins. We recently developed a new protein structure alignment program called Kpax which combines an efficient dynamic programming based scoring function with a simple but novel Gaussian representation of protein backbone shape [59]. This means that we can now quantitatively compare 3D protein domains at a similar rate to throughput to conventional protein sequence comparison algorithms. We recently compared Kpax with a large number of other structure alignment programs, and we found Kpax to be the fastest and amongst the most accurate, in a CATH family recognition test [48]. The latest version of Kpax [9] can calculate multiple flexible alignments, and thus promises to avoid such issues when comparing more distantly related protein folds and fold families.

3.1.3. Formalising and Exploiting Domain Knowledge

Concerning protein structure classification, we aim to explore novel classification paradigms to circumvent the problems encountered with existing hierarchical classifications of protein folds and domains. In particular it will be interesting to set up fuzzy clustering methods taking advantage of our previous work on gene functional classification [36], but instead using Kpax domain-domain similarity matrices. A non-trivial issue with fuzzy clustering is how to handle similarity rather than mathematical distance matrices, and how to find the optimal number of clusters, especially when using a non-Euclidean similarity measure. We will adapt the algorithms and the calculation of quality indices to the Kpax similarity measure. More fundamentally, it will be necessary to integrate this classification step in the more general process leading from data to knowledge called Knowledge Discovery in Databases (KDD) [39].

Another example where domain knowledge can be useful is during result interpretation: several sources of knowledge have to be used to explicitly characterise each cluster and to help decide its validity. Thus, it will be useful to be able to express data models, patterns, and rules in a common formalism using a defined vocabulary for concepts and relationships. Existing approaches such as the Molecular Interaction (MI) format [42] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [53]. A different point of view is represented in the Interaction Network Ontology (INO), a community-driven ontology that aims to standardise and integrate data on interaction networks and to support computer-assisted reasoning [73]. However, this ontology does not integrate basic 3D concepts and structural relationships. Therefore, extending such formalisms and symbolic relationships will be beneficial, if not essential, when classifying the 3D shapes of proteins at the domain family level.

3.1.4. 3D Protein Domain Annotation and Shape Mining

A widely used collection of protein domain families is “Pfam” [38], constructed from multiple alignments of protein sequences. Integrating domain-domain similarity measures with knowledge about domain binding sites, as introduced by us in our KBDOCK approach [1], [3], can help in selecting interesting subsets of domain pairs before clustering. Thanks to our KBDOCK and Kpax projects, we already have a rich set of tools with which we can start to process and compare all known protein structures and PPIs according to their component Pfam domains. Linking this new classification to the latest “SIFTS” (Structure Integration with Function, Taxonomy and Sequence) [67] functional annotations between standard Uniprot (<http://www.uniprot.org/>) sequence identifiers and protein structures from the Protein Data Bank (PDB) [29] could then provide a useful way to discover new structural and functional relationships which are difficult to detect in existing classification schemes such as CATH or SCOP. As part of the thesis project of Seyed Alborzi, we developed a recommender-based data mining technique to associate enzyme classification code numbers with Pfam domains using our recently developed EC-DomainMiner program [11]. We subsequently generalised this approach as a tripartite graph mining method for inferring associations between different protein annotation sources, which we call “CODAC” (for COMputational Discovery of Direct Associations using Common Neighbours). A first paper on this approach was presented at IWBBIO-2017 [23].

3.2. Integrative Multi-Component Assembly and Modeling

3.2.1. Context

At the molecular level, each PPI is embodied by a physical 3D protein-protein interface. Therefore, if the 3D structures of a pair of interacting proteins are known, it should in principle be possible for a docking algorithm to use this knowledge to predict the structure of the complex. However, modeling protein flexibility accurately during docking is very computationally expensive due to the very large number of internal degrees of freedom in each protein, associated with twisting motions around covalent bonds. Therefore, it is highly impractical to use detailed force-field or geometric representations in a brute-force docking search. Instead, most protein docking algorithms use fast heuristic methods to perform an initial rigid-body search in order to locate a relatively small number of candidate binding orientations, and these are then refined using a more expensive interaction potential or force-field model, which might also include flexible refinement using molecular dynamics (MD), for example.

3.2.2. Polar Fourier Docking Correlations

In our *Hex* protein docking program [60], the shape of a protein molecule is represented using polar Fourier series expansions of the form

$$\sigma(\underline{x}) = \sum_{nlm} a_{nlm} R_{nl}(r) y_{lm}(\theta, \phi), \quad (1)$$

where $\sigma(\underline{x})$ is a 3D shape-density function, a_{nlm} are the expansion coefficients, $R_{nl}(r)$ are orthonormal Gauss-Laguerre polynomials and $y_{lm}(\theta, \phi)$ are the real spherical harmonics. The electrostatic potential, $\phi(\underline{x})$, and charge density, $\rho(\underline{x})$, of a protein may be represented using similar expansions. Such representations allow the *in vacuo* electrostatic interaction energy between two proteins, A and B, to be calculated as [44]

$$E = \frac{1}{2} \int \phi_A(\underline{x}) \rho_B(\underline{x}) d\underline{x} + \frac{1}{2} \int \phi_B(\underline{x}) \rho_A(\underline{x}) d\underline{x}. \quad (2)$$

This equation demonstrates using the notion of *overlap* between 3D scalar quantities to give a physics-based scoring function. If the aim is to find the configuration that gives the most favourable interaction energy, then it is necessary to perform a six-dimensional search in the space of available rotational and translational degrees of freedom. By re-writing the polar Fourier expansions using complex spherical harmonics, we showed previously that fast Fourier transform (FFT) techniques may be used to accelerate the search in up to five of the six degrees of freedom [61]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [10], [6]. Consequently, we are continuing to explore new ways to exploit the polar Fourier approach.

3.2.3. Assembling Symmetrical Protein Complexes

Although protein-protein docking algorithms are improving [62], [46], it still remains challenging to produce a high resolution 3D model of a protein complex using *ab initio* techniques, mainly due to the problem of structural flexibility described above. However, with the aid of even just one simple constraint on the docking search space, the quality of docking predictions can improve considerably [10], [61]. In particular, many protein complexes involve symmetric arrangements of one or more sub-units, and the presence of symmetry may be exploited to reduce the search space considerably [28], [58], [65]. For example, using our operator notation (in which \hat{R} and \hat{T} represent 3D rotation and translation operators, respectively), we have developed an algorithm which can generate and score candidate docking orientations for monomers that assemble into cyclic (C_n) multimers using 3D integrals of the form

$$E_{AB}(y, \alpha, \beta, \gamma) = \int \left[\hat{T}(0, y, 0) \hat{R}(\alpha, \beta, \gamma) \phi_A(\underline{x}) \right] \times \left[\hat{R}(0, 0, \omega_n) \hat{T}(0, y, 0) \hat{R}(\alpha, \beta, \gamma) \rho_B(\underline{x}) \right] d\underline{x}, \quad (3)$$

where the identical monomers A and B are initially placed at the origin, and $\omega_n = 2\pi/n$ is the rotation about the principal n -fold symmetry axis. This example shows that complexes with cyclic symmetry have just 4 rigid body degrees of freedom (DOFs), compared to $6(n-1)$ DOFs for non-symmetrical n -mers. We have generalised these ideas in order to model protein complexes that crystallise into any of the naturally occurring point group symmetries (C_n , D_n , T , O , I). This approach was published in 2016 [8], and was subsequently applied to several symmetrical complexes from the ‘‘CAPRI’’ blind docking experiment [18]. Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to refine candidate solutions using a more accurate coarse-grained (CG) force-field scoring function.

3.2.4. Coarse-Grained Models

Many approaches have been proposed in the literature to take into account protein flexibility during docking. The most thorough methods rely on expensive atomistic simulations using MD. However, much of a MD trajectory is unlikely to be relevant to a docking encounter unless it is constrained to explore a putative protein-protein interface. Consequently, MD is normally only used to refine a small number of candidate rigid body docking poses. A much faster, but more approximate method is to use CG normal mode analysis (NMA) techniques to reduce the number of flexible degrees of freedom to just one or a handful of the most significant vibrational modes [52], [37], [49], [50]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [68], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [2].

In the last few years, CG *force-field* models have become increasingly popular in the MD community because they allow very large biomolecular systems to be simulated using conventional MD programs [27]. Typically, a CG force-field representation replaces the atoms in each amino acid with from 2 to 4 “pseudo-atoms”, and it assigns each pseudo-atom a small number of parameters to represent its chemo-physical properties. By directly attacking the quadratic nature of pair-wise energy functions, coarse-graining can speed up MD simulations by up to three orders of magnitude. Nonetheless, such CG models can still produce useful models of very large multi-component assemblies [64]. Furthermore, this kind of coarse-graining effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [43]. We are therefore developing a “coarse-grained” scoring function for fast protein-protein docking and multi-component assembly in the frame of the PhD project of Maria-Elisa Ruiz-Echartea.

3.2.5. Assembling Multi-Component Complexes and Integrative Structure Modeling

We also want to develop related approaches for integrative structure modeling using cryo-electron microscopy (cryo-EM). Thanks to recently developments in cryo-EM instruments and technologies, it is now feasible to capture low resolution images of very large macromolecular machines. However, while such developments offer the intriguing prospect of being able to trap biological systems in unprecedented levels of detail, there will also come an increasing need to analyse, annotate, and interpret the enormous volumes of data that will soon flow from the latest instruments. In particular, a new challenge that is emerging is how to fit previously solved high resolution protein structures into low resolution cryo-EM density maps. However, the problem here is that large molecular machines will have multiple sub-components, some of which will be unknown, and many of which will fit each part of the map almost equally well. Thus, the general problem of building high resolution 3D models from cryo-EM data is like building a complex 3D jigsaw puzzle in which several pieces may be unknown or missing, and none of which will fit perfectly. We wish to proceed firstly by putting more emphasis on the single-body terms in the scoring function [35], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space.

DYLISS Project-Team

3. Research Program

3.1. Modeling knowledge integration with combinatorial constraints

Biological networks are built with data-driven approaches aiming at translating genomic information into a functional map. Most methods are based on a probabilistic framework which defines a probability distribution over the set of models. The reconstructed network is then defined as the most likely model given the data.

Our team has investigated an alternative perspective where each data induces a set of constraints – related to the steady state response of the system dynamics – on the set of possible values in a network of fixed topology. The methods that we have developed complete the network with product states at the level of nodes and influence types at the level of edges, able to globally explain experimental data. In other words, the selection of relevant information in the model is no more performed by selecting *the* network with the highest score, but rather by exploring the complete space of models satisfying constraints on the possible dynamics supported by prior knowledge and observations. In the (common) case when there is no model satisfying all the constraints, we relax the problem by introducing new combinatorial optimization problems that introduce the possibility of correcting the data or the knowledge. Common properties to all solutions are considered as a robust information about the system, as they are independent from the choice of a single solution to the optimization problem [77].

Solving these computational issues requires addressing NP-hard qualitative (non-temporal) issues. We have developed a long-term collaboration with Potsdam University in order to use a logical paradigm named **Answer Set Programming** (ASP) [46], [73] to solve these constraint satisfiability and combinatorial optimization issues. Applied on transcriptomic or cancer networks, our methods identified which regions of a large-scale network shall be corrected [47], and proposed robust corrections [6]. This result suggested that this approach was compatible with efficiency, scale and expressivity needed by biological systems.

During the last years, our goal was to provide **formal models of queries on biological networks** with the focus of integrating dynamical information as explicit logical constraints in the modeling process. Using these technologies requires to revisit and reformulate constraint-satisfiability problems at hand in order both to decrease the search space size in the grounding part of the process and to improve the exploration of this search space in the solving part of the process. Concretely, getting logical encoding for the optimization problems forces to clarify the roles and dependencies between parameters involved in the problem. This paves the way to a refinement approach based on a fine investigation of the space of hypotheses in order to make it smaller and gain in the understanding of the system. Our studies confirmed that logical paradigms are a powerful approach to build and query reconstructed biological systems, in complement to discriminative ("black-box") approaches based on statistical machine-learning. Based on these technologies, we have developed a panel of methods allowing the integration of multi-scale data knowledge, linking genomics, metabolomics, expression data and protein measurement of several phenotypes.

Notice that our main issue is in the field of knowledge representation. More precisely, we do not wish to develop new solvers or grounders, a self-contained computational issue which is addressed by specialized teams such as our collaborator team in Potsdam. Our goal is rather to investigate how the constant progresses in the field of constraint logical programming, shown by the performance of ASP-solvers, are sufficient to address the complexity of constraint-satisfiability and combinatorial optimization issues explored in systems biology. In this direction, we work in close interaction with Potsdam university to feed their research activities with challenging issues from bioinformatics and, as a feed-back, take benefit of the prototypes they develop.

By exploring the complete space of models, our approach typically produces numerous candidate models compatible with the observations. We began investigating to what extent domain knowledge can further refine the analysis of the set of models by identifying classes of similar models, or by selecting a subset of models that satisfy an additional constraint (for instance, best fit with a set of experiments, or with a minimal size). We anticipate that this will be particularly relevant when studying non-model species for which little is known but valuable information from other species can be transposed or adapted. These efforts consist in developing reasoning methods based on ontologies as formal representation of symbolic knowledge. We use Semantic Web tools such as SPARQL for querying and integrating large sources of external knowledge, and measures of semantic similarity and particularity for analyzing data.

3.2. Modeling the dynamical response of biological systems with logical and (non)-linear constraints

As explained below, Answer Set Programming technologies enable the identification of key controllers based on the integration of static data. As a natural follow-up, we also develop optimization techniques to learn models of the dynamics of a biological system. As before, our strategy is not to select a single model fitting with experimental data but rather to decipher the complete set of families of models which are compatible with the observed response. Our main research line in this field is to decipher the appropriate level of expressivity (in terms of constraints) allowing both to properly report the nature of data and knowledge and to allow for an exhaustive study of the space of feasible models. To implement this strategy, we rely on several constraint programming frameworks, which depend on the model scale and the nature of time-points kinetic measurements. Logical programming (Answer Set Programming) is used to decipher the combinatorics of synchronic Boolean networks explaining static or dynamics response of signaling networks to perturbations (such as measured by phosphoproteomics technologies) [7]. SAT-based approaches are used to decipher the combinatorics of large-scale asynchronous boolean networks. In order to gain in expressivity, we model these networks as guarded-transition network, an extension of Petri nets [40]. Finally, classical learning methods are used to build ad-hoc parameterized numerical models that provide the most parsimonious explanations to experimental measurements.

3.3. Modeling sequences with formal grammars

Once groups of genome products involved in the answer of the species have been identified with integrative or dynamical methods, it remains to characterize the biological actors within genomes. To that goal, we both learn, model and parse formal patterns within DNA, RNA or protein sequences. More precisely, our research on modeling biomolecular sequences with expressive formal grammars focuses on learning such grammars from examples, helping biologists to design their own grammar and providing practical parsing tools.

On the development of **machine learning** algorithms for the induction of grammatical models [66], we have a strong expertise on learning finite state automata. We have proposed an algorithm that learns successfully automata modeling families of (non homologous) functional families of proteins [5], leading to a tool named Protomata-learner. The algorithm is based on a similar fragment merging heuristic approach which reports partial and local alignments contained in a family of sequences. As an example, this tool allowed us to properly model the TNF protein family, a difficult task for classical probabilistic-based approaches. It was also applied successfully to model important enzymatic families of proteins in cyanobacteria [4]. Our future goal is to further demonstrate the relevance of formal language modeling by addressing the question of a fully automatic prediction from the sequence of all the enzymatic families, aiming at improving even more the sensitivity and specificity of the models. As enzyme-substrate interactions are very specific central relations for integrated genome/metabolome studies and are characterized by faint signatures, we shall rely on models for active sites involved in cellular regulation or catalysis mechanisms. This requires to build models gathering both structural and sequence information in order to describe (potentially nested or crossing) long-term dependencies such as contacts of amino-acids that are far in the sequence but close in the 3D protein folding. Our current researches is focused on the inference of Context-Free Grammars including the topological information coming from the structural characterization of active sites.

Using context-free grammars instead of regular patterns increases the complexity of **parsing** issues. Indeed, efficient parsing tools have been developed to identify patterns within genomes but most of them are restricted to simple regular patterns. Definite Clause Grammars (DCG), a particular form of logical context-free grammars have been used in various works to model DNA sequence features [86]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search [110], [109]. This increases the expressivity of the formalism towards mildly context sensitive grammars. Thus, those grammars model not only DNA/RNA sequence features but also structural features such as repeats, palindromes, stem/loop or pseudo-knots. A few years ago, we have designed a first tool, STAN (suffix-tree analyser), in order to make it possible to search for a subset of SVG patterns in full chromosome sequences [92]. This tool was used for the recognition of transposable elements in *Arabidopsis thaliana* [113]. We have enlarged this experience through a new modeling language, called Logol [1]. Generally, a suitable language for the search of particular components in languages has to meet several needs : expressing existing structures in a compact way, using existing databases of motifs, helping the description of interacting components. In other words, the difficulty is to find a good tradeoff between expressivity and complexity to allow the specification of realistic models at genome scale. The Logol language and associated framework have been built in this direction. The Logol specificity besides other SVG-like languages mainly lies in a systematic introduction of constraints on string variables.

3.4. Symbolic methods for model space exploration: Semantic web for life sciences and Formal Concepts Analysis

All the methods presented in the previous sections usually result in pools of candidates which equivalently explain the data and knowledge. These candidates can be dynamical systems, compounds, biological sequences, proteins... In any case, the output of our formal methods generally requires *a posteriori* investigation and filtering by domain experts. In order to assist them, we rely on two classes of symbolic technics: Semantic Web technologies and Formal Concept Analysis (FCA). They both aim at the formalization and management of knowledge, that is, the explicitation of relations occurring in structured data. These technics complement each other: the production of relevant concepts in FCA highly depends on the availability of semantic annotations using a controlled set of terms and conversely, building and exploiting ontologies is a complex process that can be made much easier with FCA.

Integrating heterogenous data with semantic web technologies The emergence of ontologies in biomedical informatics and bioinformatics happened in parallel with the development of the **Semantic Web** in the computer science community [108]. Let us recall that the Semantic Web is an extension of the current Web that provides an infrastructure integrating data and ontologies in order to support unified reasoning. Since the beginning, life sciences have been a major application domain for the Semantic Web [48]. This was motivated by the joint evolution of data acquisition capabilities in the biomedical field, and of the methods and infrastructures supporting data analysis (grids, the Internet...), resulting in an explosion of data production in complementary domains [58], [49]. Consequently, Semantic Web technologies have become an integral part of translational medicine and translational bioinformatics [63]. The Linked Open Data project promotes the integration of data sources in machine-processable formats compatible with the Semantic Web [57], with a strong involvement of life sciences in this initiative.

However, a specificity of life sciences “data deluge” is that the proportion of generated data is much higher than in the more general “big data phenomenon”, and that these data are highly connected [112]. **The bottleneck that once was data scarcity now lies in the lack of adequate methods supporting data integration, processing and analysis** [88]. Each of these steps typically hinges on domain knowledge, which is why they resist automation. This knowledge can be seen as the set of rules representing in what conditions data can be used or can be combined for inferring new data or new links between data.

In this setting, we are working on the integration of Semantic Web resources with our data analysis methods in order to take existing biological knowledge into account. We have introduced several methods to interpret semantic similarities and particularities [56], [55]. We now focus our attention on the semi-automated construction of RDF abstractions of heterogeneous datasets which can be handled by non-expert users. This

allows both to automatically prepare input datasets for the other methods developed in the team and to analyse the output of the methods in a wide knowledge context.

Using Formal concept analysis to explore the results of bioinformatics analyses Formal concept analysis aims at the development of conceptual structures which can be logically activated for the formation of judgments and conclusions [117]. It is used in various domains managing structured data such as knowledge processing, information retrieval or classification [90]. In its most simple form, one considers a binary relation between a set of objects and a set of attributes. In this setting, formal concept analysis formalizes the semantic notions of extension and intension. Concepts are related within a lattice structure (Galois connection) by subconcept-superconcept relations, and this allows drawing causality relations between attribute subsets. In bioinformatics, it has been used to derive phylogenetic relations among groups of organisms [87], a classification task that requires to take into account many-valued Galois connections. We have proposed in a similar way a classification scheme for the problem of protein assignment in a set of protein families [67].

One of the most important issue with concept analysis is due to the fact that current methods remain very sensitive to the presence of uncertainty or incompleteness in data. On the other hand, this apparent defect can be reversed to serve as a marker of incompleteness or inconsistency [68]. Following this inspiration, we have proposed a methodology to tackle the problem of uncertainty on biological networks where edges are mostly predicted links with a high level of false positives [118]. The general idea consists to look for a tradeoff between the simplicity of the conceptual representation and the need to manage exceptions. As a very prospective challenge, we are exploring the idea of using ontologies to help this or to help ontology refinement using concept analysis [93], [52], [102].

More generally, common difficult tasks in this context are visualization, search for local structures (graph mining) and network comparison. Network compression is a good solution for an efficient treatment of all these tasks. This has been used with success in power graphs, which are abstract graphs where nodes are clusters of nodes in the initial graph and edges represent bicliques between two sets of nodes [104]. In fact, concepts are maximal bicliques and we are currently developing the power graph idea in the framework of concept analysis.

3.5. Implementing methods in software and platforms

Seven platforms have been developped in the team for the last five years: Askomics, AuReMe, FinGoc, Caspo, Cadiom, Logol, Protomata. Indeed, one of the team's goals is to facilitate interplays between the tools for biological data analysis and integration. Improvements and novelties of these platforms are described in the "software" section. Our platforms aim at guiding the user to progressively reduce the space of models (families of sequences of genes or proteins, families of keys actors involved in a system response, dynamical models) which are compatible with both knowledge and experimental observations.

Most of our platforms are developed with the support of the GenOuest resource and data center hosted in the IRISA laboratory, including their computer facilities [more info]. It worths considering them into larger dedicated environments to benefit from the expertise of other research groups. The **BioShaddock** repository of the GenOuest platform allows one to share the different docker containers that we are developing [website]. The **GenOuest galaxy portal** of the GenOuest platform now provides access to most tools for integrative biology and sequence annotation (access on demand).

3.5.1. AskOmics platform

Goal Integration and interrogation software for linked biological data based on semantic web technologies [url].

Description *AskOmics* aims at bridging the gap between end user data and the Linked (Open) Data cloud. It allows heterogeneous bioinformatics data (formatted as tabular files or directly in RDF) to be loaded into a Triple Store system using a user-friendly web interface. *AskOmics* also provides an intuitive graph-based user interface supporting the creation of complex queries that currently require hours of manual searches across tens of spreadsheet files. The elements of interest selected in the graph are then automatically converted into a SPARQL query that is executed on the users's data.

Originality Our experience is that end users (i) do not benefit for all the information available in the LOD cloud repositories by lack of SPARQL expertise (understandably: they are biologists and most of them do not have an interest in either learning SPARQL nor in learning how to integrate data); (ii) do not contribute their data back to the LOD cloud. Again, they do not have the expertise nor the resources to produce and maintain datasets and the associated metadata as linked data, nor to maintain the underlying server infrastructure. Therefore there is a need for helping end users to (1) take advantage of the information readily available in the LOD cloud for analyzing their own data and (2) contribute back to the linked data by representing their data and the associated metadata in the proper format as well as by linking them to other resources. In this context, the main originality is the graphical interface that allows any SPARQL query to be built transparently and iteratively by a non-expert user.

Application This software was developed in the context of the MirnAdapt (pea-aphid) project in 2016. The tool has been presented to the agriculture communities in conferences [53], [84] and to the Galaxy community [29]. Up to now, more than 10 biological partners team are actually testing and using the prototype software (colza, pea-aphids, copper microbiology, marine biology), and SANOFI has shown its interest to co-develop the tool. Even if its current user base belongs to the bioinformatics community, the scope of *AskOmics* is domain-independent and has the potential to reach a wider audience related to the Semantic Web community.

3.5.2. *AuReMe workspace*

Goal Tractable reconstruction of metabolic networks [url].

Description The toolbox *AuReMe* allows for the Automatic Reconstruction of Metabolic networks based on the combination of multiple heterogeneous data and knowledge sources [64]. It is available as a Docker image. Five modules are composing *AuReMe*: 1) The Model-management PADmet module allows manipulating and tracing all metabolic data via a local database. [package] 2) The meneco python package allows the gaps of a metabolic network to be filled by using a topological approach that implements a logical programming approach to solve a combinatorial problem [107], [65] and [21] [python package] 3) The shogen python package allows genome and metabolic network to be aligned in order to identify genome units which contain a large density of genes coding for enzymes; it also implements a logical programming approach [60] [python package]. 4) The manual curation assistance PADmet module allows the reported metabolic networks and their metadata to be curated. 5) The Wiki-export PADmet module enables the export of the metabolic network and its functional genomic unit as a local wiki platform allowing a user-friendly investigation [package].

Originality The main added-values are the inclusion of graph-based gap-filling tools that are particularly relevant for the study of non-classical organisms, the possibility to trace the reconstruction and curation procedures, and the representation and exploration of reconstructed metabolic networks with wikis.

Application The tools included in *AuReMe* have been used for reconstructing metabolic networks of micro and macro-algae [97], extremophile bacteria [16] and communities of organisms [61] in the context of the Idealg, Ciric-omics and IPL Algae-In-Silico projects.

3.5.3. *FinGoc-tools*

Goal Filtering interaction networks with graph-based optimization criteria.

Description The goal is to offer a set of tools for the reconstruction of networks from genome, literature and large-scale observation data (expression data, metabolomics...) in order to elucidate the main regulators of an observed phenotype. Most of the optimization issues are addressed with Answer Set Programming. 1) The **lombarde package** enables the filtering of transcription-factor/binding-site regulatory networks with mutual information reported by the response to environmental perturbations. The high level of false-positive interactions is filtered according to graph-based criteria. Knowledge about regulatory modules such as operons or the output of the shogen package can be taken into account [39], [38] [web server]. 2) The **KeyRegulatorFinder package** allows searching key regulators of lists of molecules (like metabolites, enzymes or genes) by taking advantage of knowledge databases in cell metabolism and signaling. The complete information is transcribed into a large-scale interaction graph which is filtered to report the most significant upstream regulators of the considered list of molecules [59] [package]. 3) The **powerGrasp**

python package provides an implementation of graph compression methods oriented toward visualization, and based on power graph analysis. [package]. 4) The **iggy package** enables the repairing of an interaction graph with respect to expression data. It proposes a range of different operations for altering experimental data and/or a biological network in order to re-establish their mutual consistency, an indispensable prerequisite for automated prediction. For accomplishing repair and prediction, we take advantage of the distinguished modeling and reasoning capacities of Answer Set Programming. [6] [114] [Python package]

Originality The main added-value of these tools is to make explicit the criteria used to highlight the role of the main regulators: the underlying methods encode explicit graph-based criteria instead of relying on statistical approaches. This makes it possible to explain local relationships and patterns within interaction graphs by explicit biological relationships.

Application The tools have been used to figure out the main gene-regulators of the response of porks to several diets in [74], [76] and [18]. The tools were also used to decipher regulators of reproduction for the pea aphid, an insect that is a pest on plants [85], [119].

3.5.4. *Caspo* software

Participant: Anne Siegel.

Goal Studying synchronous boolean networks [url]

Description Cell ASP Optimizer (*Caspo*) constitutes a pipeline for automated reasoning on logical signaling networks. The main underlying issue is that inherent experimental noise is considered, many different logical networks can be compatible with a set of experimental observations (see [106] and [22]). It is available as a Docker container. Five modules are composing *Caspo*: 1) the *Caspo-learn* module performs an automated inference of logical networks from experimental data allows for identifying admissible large-scale logic models saving a lot of efforts and without any a priori bias [115] and [78]. 2) The *Caspo-classify*, predict and visualize modules allows for classifying a family of boolean networks with respect to their input-output predictions [78]. 3) The *Caspo-design* module designs experimental perturbations which would allow for an optimal discrimination of rival models in a family of boolean networks [116]. 4) The *Caspo-control* module identifies key-players of a family of networks: it computes robust intervention strategies that force a set of target species or compounds into a desired steady state [80]. 5) The *Caspo-timeseries* module to take into account time-series observation datasets in the learning procedure [94] [python package and docker container].

Originality The *Caspo* modules provide friendly and efficient solutions to problems that were previously addressed in theoretical papers with MILP programs. The main advantage is that it enables a complete study of logical network without requiring any linear constraint programs.

Application The *Caspo* tool was initiated in the framework of the BioTempo project. *Caspo-learn* has been included as a module to learn logical networks from early steady-state data in CellNopt, a generic platform which implements several methods for learning and studying signaling networks at different modeling levels (from logical models to numerical models).

3.5.5. *Cadbiom* package

Goal Building and analyzing the asynchronous dynamics of enriched logical networks [url]

Description Based on Guarded transition semantic, the *Cadbiom* software provides a formal framework to help the modeling of biological systems such as cell signaling network. It allows synchronization events to be investigated in biological networks [40]. It is available as a Docker image. Three modules are composing *Cadbiom*: 1) The *Cadbiom* graphical interface is useful to build and study moderate size models. It provides exploration, simulation and checking. For large-scale models, *Cadbiom* also allows to focus on specific nodes of interest. 2) The *Cadbiom* API allows a model to be loaded, performing static analysis and checking temporal properties on a finite horizon in the future or in the past. 3) Exploring large-scale knowledge repositories, the translations of the large-scale PID repository (about 10,000 curated interactions) have been translated into the *Cadbiom* formalism.

Originality Model-checking approaches applied to Boolean networks [81] or multivalued networks [91] allow the trajectories of the system to be entirely studied but they can only be applied to small-size networks. On the contrary, *Cadbiom* is able to handle large-scale knowledge databases.

Application The *Cadbiom* tool was applied to study the regulators of the TGF- β , a gene that controls liver fibrosis [40] in the framework of the *TGFSysBio* project. The study of its predictions also enabled large-scale knowledge databases (PID) to be curated [25].

3.5.6. *Logol* software

Goal Complex pattern modelling and matching [url]

Description The *Logol* toolbox is a swiss-army-knife for pattern matching on DNA/RNA/Protein sequences, using a high-level grammatical formalism to permit a large expressivity for patterns [50]. A *Logol* pattern can consist in a complex combination of motifs (such as degenerated strings) and structures (such as imperfect stem-loop or repeats). *Logol* key features are the possibilities to divide a pattern description into several sub-patterns, to model long range dependencies, to enable the use of ambiguous models or to permit the inclusion of negative conditions in a pattern definition. The *LogolMatch* parser takes as input a biological sequence and a grammar file. It returns a XML file containing all the occurrences of the pattern in the sequence with their parsing details. The input sequences can be genomes from biological banks.

Originality Many pattern matching tools exist to efficiently model specific types of patterns: *vmatch* [82], *patmatch* [121], *cutadapt* [89], scoring matrix or profile HMMs [44], [71]. The main advantage of *Logol* is its very large expressivity. It encompasses most of the features of these specialized tools and enables interplays between several classes of patterns (motifs and structures).

Application The *Logol* tool was applied to the detection of mutated primers in a metabarcoding study [41], [42] or to stem-loop identification (e.g. in CRISPR ⁰ [103], [50]). Ongoing application is the search for transposable elements in the human genome in the context of a colorectal cancer study ⁰. *Logol* strongly supported the study of the LXR- α targets in the framework of the *FatInteger* project.

3.5.7. *Protomata-suite*

Goal Expressive pattern discovery on protein sequences [url]

Description *Protomata* is a machine learning suite for the inference of automata characterizing (functional) families of proteins from available sequences. Based on partial and local alignments, *Protomata* learns precise characterizations of the families of proteins, allowing new family members to be predicted with a high specificity. Three main modules are integrated in the *Protomata-learner* workflow are available as well as stand-alone programs: 1) *paloma* builds partial local multiple alignments, 2) *protobuild* infers automata from these alignments and 3) *protomatch* and *protoalign* scans, parses and aligns new sequences with learnt automata. The suite is completed by tools to handle or visualize data and can be used online by the biologists via a web interface on Genouest Platform. It is actively maintained (version v2.1 was released in April 2017) and we are scheduling a new major version with enhanced scoring schemes that we have proposed [105].

Originality The main specificity is that the power of characterization is beyond the scope of classical sequence patterns such as PSSM (e.g. MEME suite [43]), Profile HMM (e.g. HMMER package [71]), or Prosite Patterns [111] allowing new family members to be predicted with a high specificity.

Application The *Protomata* tool is used both to update automatically the Cyanolase database [62] and, when combined to Formal Concept Analysis, to automated enzyme classification, such as the HAD superfamily of proteins [67] in the framework of the *Idealg* project.

⁰<http://crispi.genouest.org/>, <https://hal.inria.fr/hal-00643408>

⁰Preprint: <http://www.biorxiv.org/content/early/2017/03/09/115030>

ERABLE Project-Team

3. Research Program

3.1. Two main goals

ERABLE has two main goals, one related to biology and the other to methodology (algorithms, combinatorics, statistics). In relation to biology, the main goal of ERABLE is to contribute, through the use of mathematical models and algorithms, to a better understanding of close and often persistent interactions between “collections of genetically identical or distinct self-replicating cells” which will correspond to organisms/species or to actual cells. The first will cover the case of what has been called symbiosis, meaning when the interaction involves different species, while the second will cover the case of a (cancerous) tumour which may be seen as a collection of cells which suddenly disrupts its interaction with the other (collections of) cells in an organism by starting to grow uncontrollably.

Such interactions are being explored initially at the molecular level. Although we rely as much as possible on already available data, we intend to also continue contributing to the identification and analysis of the main genomic and systemic (regulatory, metabolic, signalling) elements involved or impacted by an interaction, and how they are impacted. We started going to the populational and ecological levels by modelling and analysing the way such interactions influence, and are or can be influenced by the ecosystem of which the “collections of cells” are a part. The key steps are:

- identifying the molecular elements based on so-called omics data (genomics, transcriptomics, metabolomics, proteomics, etc.): such elements may be gene/proteins, genetic variations, (DNA/RNA/protein) binding sites, (small and long non coding) RNAs, etc.
- simultaneously inferring and analysing the network that models how these molecular elements are physically and functionally linked together for a given goal, or find themselves associated in a response to some change in the environment;
- modelling and analysing the populational and ecological network formed by the “collections of cells in interaction”, meaning modelling a network of networks (previously inferred or as already available in the literature);
- analysing how the behaviour and dynamics of such a network of networks might be controlled by modifying it, including by subtracting some of its components from the network or by adding new ones.

In relation to methodology, the main goal is to provide those enabling to address our main biological objective as stated above that lead to the best possible interpretation of the results within a given pre-established model and a well defined question. Ideally, given such a model and question, the method is exact and also exhaustive if more than one answer is possible. Three aspects are thus involved here: establishing the model within which questions can and will be put; clearly defining such questions; exactly answering to them or providing some guarantee on the proximity of the answer given to the “correct” one. We intend to continue contributing to these three aspects:

- at the modelling level, by exploring better models that at a same time are richer in terms of the information they contain (as an example, in the case of metabolism, using hypergraphs as models for it instead of graphs) and are susceptible to an easier treatment:
 - these two objectives (rich models that are at the same time easy to treat) might in many cases be contradictory and our intention is then to contribute to a fuller characterisation of the frontiers between the two;
 - even when feasible, the richer models may lack a full formal characterisation (this is for instance the case of hypergraphs) and our intention is then to contribute to such a characterisation;

- at the question level, by providing clear formalisations of those that will be raised by our biological concerns;
- at the answer level:
 - to extend the area of application of exact algorithms by: (i) a better exploration of the combinatorial properties of the models, (ii) the development of more efficient data structures, (iii) a smarter traversal of the space of solutions when more than one solution exists;
 - when exact algorithms are not possible, or when there is uncertainty in the input data to an algorithm, to improve the quality of the results given by a deeper exploration of the links between different algorithmic approaches: combinatorial, randomised, stochastic.

3.2. Different research axes

The goals of the team are biological and methodological, the two being intrinsically linked. Any division into axes along one or the other aspect or a combination of both is thus somewhat artificial. Our choice is based more on the biological questions as these are a main (but not unique) driver for the methodological developments. However, since another main objective is to contribute to the fields of exact enumeration algorithms and of combinatorics, we also defined an axis that is exclusively oriented towards some of the more theoretical aspects of such objective in as much as these can be abstracted from the biological motivation. This will concern improving theory and deeply exploring the links between different algorithmic approaches: combinatorial, randomised, stochastic.

Initially, when ERABLE was created, five axes were defined. The first four fell in the first category above, and the fifth one in the second.

More recently however, as was indicated in the evaluation report for the period 2013-2017, a new biological axis was added to the four that existed already. This axis is specifically oriented towards health in general, human or animal. It was numbered as Axis 4 as the last biological one which existed already may be seen as a generalisation of the first four (three old ones and new fourth). Indeed, one overall objective of ERABLE for the next four years will be to try to establish the links between non infectious diseases such as cancer or rare ones on one hand, and infectious diseases (related to symbiosis understood in its more general sense) on the other.

As concerns symbiosis, the model organisms or systems chosen include the following cases:

- Arthropods, notably insects, and their parasites;
- Symbiont-harboring trypanosomatids and trypanosomes more in general;
- The bacterial communities inside the respiratory tract of mammals (swine, bovine);
- Human in general, and the human microbiota in particular also for its possible relation to cancer.

Notice however that: (1) new model organisms or systems may be considered as the opportunity for new collaborations appears, indeed such collaborations will be actively searched for; and (2) we will always attempt to explore mathematical and computational models and to develop algorithmic methods that are as much as possible generic.

Axis 1: Identifying the molecular elements

Intra and inter-cellular interactions involve molecular elements whose identification is crucial to understand what governs, and also what might enable to control such interactions. For the sake of clarity, the elements may be classified in two main classes, one corresponding to the elements that allow the interactions to happen by moving around or across the cells, and another that are the genomic regions where contact is established. Examples of the first are non coding RNAs, proteins, and mobile genetic elements such as (DNA) transposons, retro-transposons, insertion sequences, etc. Examples of the second are DNA/RNA/protein binding sites and targets. Furthermore, both types (effectors and targets) are subject to variation across individuals of a population, or even within a single (diploid) individual. Identification of these variations is yet another topic that we wish to cover. Variations are understood in the broad sense and cover single nucleotide polymorphisms

(SNPs), copy-number variants (CNVs), repeats other than mobile elements, genomic rearrangements (deletions, duplications, insertions, inversions, translocations) and alternative splicings (ASs). All three classes of identification problems (effectors, targets, variations) may be put under the general umbrella of genomic functional annotation.

Axis 2: Inferring and analysing the networks of molecular elements

As increasingly more data about the interaction of molecular elements (among which those described above) becomes available, these should then be modelled in a subsequent step in the form of genetic, metabolic, protein-protein interaction and signalling networks. This raises two main classes of problems. The first is to accurately infer such networks. Reconstructing, by analogy, the metabolic network of an organism is often considered, rightly or wrongly, to be easier than inferring a gene regulatory network, also because in the latter case, identifying all the elements participating in the network is in itself a complex and far from solved issue, as we saw in Axis 1. Moreover, the difficulty varies depending on whether only the structure or also the dynamics of the network is of interest, assuming that the latter may be studied (kinetics data are often missing even with the increasingly more sophisticated and performing technologies we have nowadays). A more complete picture of the functioning of a cell would further require that ever more layers of network and molecular profile data, when available, are integrated together, which raises the problem of how to model together information that is heterogeneous at different levels. Modelling together metabolic and gene regulation for instance is already a hard problem given that the two happen at very different time-scales: fast for metabolic regulation, slow for gene regulation.

Even assuming such a network, integrated or “simple”, has been inferred for a given organism or set of organisms, the second problem is then to develop the appropriate mathematical models and methods to extract further biological information from such networks. The difficulty of this differs of course again depending on whether only the structure of the network is of interest, or also its dynamics. We are addressing various questions related to one or the other of the above aspects – inference and analysis.

Axis 3: Modelling and analysing a network of individuals, or a network of individuals’ networks

As mentioned, at its extreme, life can be seen as one collection, or a collection of collections of genetically identical or distinct self-replicating cells who interact, sometimes closely and for long periods of evolutionary time, with a same or with distinct functional objectives. One striking example is human, who is composed of cells which are both native and extraneous; in fact, a surprising 90% is believed to belong to the second category, mostly bacteria, including one which lost its identity to become a “mere” human organelle, the mitochondrion. Bacteria on the other hand group into colonies of genetically identical individuals which may sometimes acquire the ability to become specialised for different tasks. Which is the “individual”, a single bacterium or a group thereof is difficult to say. To understand human or bacteria, or to understand any other organism, it appears therefore essential to better comprehend the interactions in which they are involved. Methodologically speaking, we must therefore move towards modelling and analysing not a single individual anymore but a network of individuals. Ultimately, we should move towards investigating a network of individuals’ networks. Moreover, since organisms interact not only with others but also with their abiotic environment, there is a need to model full ecosystems, at a static but also at a dynamic level, that is by taking into account the fact that individuals or populations move in space. Our intention at a longer term is to address all such different levels. We started with the molecular and static one that we are treating from different perspectives for a large number of species at the genomic level (Baudet *et al.*, Syst Biol, 64(3), 2015) and for a small number at the network level (Cottret *et al.*, PLoS Comput Biol, 6(9), 2010). We intend in a near future to slowly move towards a populational and ecological approach that is dynamic in both time and space.

Axis 4: Human and animal health

As indicated above, this is a recent axis in the team and concerns various applications to human and animal health. In some ways, it overlaps with the three previous axes as well as with Axis 5 on the methodological aspects, but since it gained more importance in the past few years, we decided to develop more these particular applications. Most of them started through collaborations with clinicians. Such applications are currently focused on three different topics: (i) Infectiology, (ii) Rare diseases, and (iii) Cancer.

Infectiology is the oldest one and is covered also by Axis 5 below. It will thus be described there. It started by a collaboration with Arnaldo Zaha from the Federal University of Rio Grande do Sul in Brazil. Rare Diseases on the other hand started by a collaboration with clinicians from the Centre de Recherche en Neurosciences of Lyon (CNRL) and is focused the Taybi-Linder Syndrome (TALS) and on abnormal splicing of U12 introns, while Cancer rests on a collaboration with the Centre Léon Bérard (CLB) and Centre de Recherche en Cancérologie of Lyon (CRCL) which is focused on Breast and Prostate carcinomas and Gynaecological carcinosarcomas.

The latter collaboration was initiated through a relationship between a member of ERABLE (Alain Viari) and Dr. Gilles Thomas who had been friends since many years. G. Thomas was one of the pioneers of Cancer Genomics in France. After his death in 2014, Alain Viari took the (part time) responsibility of his team at CLB and pursued the main projects he had started.

Within Inria and beyond, the first two applications (Infectiology and Rare Diseases) may be seen as unique because of their specific focus (resp. respiratory tract of swines and TALS). In the first case, such uniqueness is also related to the fact that the work done involves a strong computational part but also experiments *performed within ERABLE itself*.

Axis 5: Going towards control

What was described in the Axes 2 and 3 above concerned modelling and analysing a molecular network, or network of networks, but not attempting to control the network at either level for bio-technological, environmental or health purposes.

In the bio-technological case, the objective can be briefly described as involving the manipulation of a species, in general a bacterium, in order for it to produce more of a given chemical compound it already synthesises (for instance, ethanol) but not in enough quantity, or to produce a metabolite it normally is not able to synthesise. The motivation for transplanting its production in a bacterium is, again, to be able to make it more effective.

As concerns control for environmental or health purposes, this could be achieved at least in some cases by manipulating the symbionts with which an organism, insect pest for instance, or humans leave. In the environmental case, this has gone under the name of “biological control” (see for instance Flint & Dreistat, “Natural Enemies Handbook: The Illustrated Guide to Biological Pest Control”, University of California Press, 1998) and involves the use of “natural enemies” of a pest organism. This idea has a long history: the ancient Chinese, observing that ants were effective predators of many citrus pests, decided to increase the ants population by displacing their nests from the surrounding habitats and placing them inside their orchards to protect them. More recently, there has been growing evidence that some endosymbiotic bacteria, that is bacteria that live within the cells of their hosts, could become efficient biocontrol agents. This is in particular the case of *Wolbachia*, a bacterium much studied in ERABLE (Ahantarig & Kittayapong, J Appl Entomology, 135(7):479-486, 2011).

The connection between disease and the disruption of homeostatic interactions between the host and its microbiota is on the other hand now well established. Microbiota-targeted therapies involve altering the community composition by eliminating individual strains of a single species (for example, with antibiotics) or replacing the entire community with a new intact microbiota. Secondary infections linked to antibiotic use provide however a cautionary tale of the possible consequences of perturbing a microbial species network.

Besides the biotechnological aspects on which we are already working in the context of two European projects (BacHBerry, and to a lesser extent, MicroWine), our main goal in this case is to try to formalise such type of control. There are two objectives here. One is methodological and concerns attempting to provide a single formal framework for the diverse ways of controlling a network, or a network of networks. Our attention has concentrated initially on metabolism, and will at a mid to longer term include regulation. Our intention notably as concerns the incorporation of regulation is to collaborate with other Inria teams, most notably IBIS with whom we are already in discussion. The second objective is biological and concerns control for environmental and health purposes. The originality we are seeking in this case is to attempt such control not by eliminating species, which is done mainly through the use of antibiotics that may then create resistance, a phenomenon that is becoming a major clinical and public health problem, but by manipulating the species or their environment,

or by changing the composition of the community by adding or displacing some other species in such a way that new equilibria may be reached which enable all the species living in a same niche to survive. The idea is not new: the areas of prebiotics (non-digestible food ingredients that stimulate the growth and/or activity of bacteria in the digestive system in beneficial ways) and probiotics (micro-organisms claimed to provide benefits when consumed) indeed cover similar concerns in relation to health. Other novel approaches propose to work at the level of bacterial communication (quorum sensing) to control for pathogenicity (Rutherford & Bassler, Cold Spring Harbor Perspectives in Medicine, 2012). Small RNAs in particular are believed to play an important role in quorum sensing.

Axis 6: Cross-fertilising different computational approaches

In computer science and in optimisation, different approaches and techniques have been proposed to cope with hardness results. It is clear that none of them is dominant: there are classes of problems for which approach A is better than approach B, and vice-versa. Moreover, there is no satisfactory understanding of the conditions that favour one approach with respect to another one.

As an example, the team that gave birth to ERABLE, BAMBOO, had expertise more in the area of combinatorial algorithms for strings (sequences), trees and graphs. Many such algorithms addressed an enumeration problem: given a certain description of the object(s) searched for or definition of a function to be optimised, the method was supposed to list all the solutions. In many real life situations, notably in biology, a majority of the problems treated, of whatever kind, enumeration or else, are however hard. Although combinatorics remains crucial to better understand the structure of such problems and delimit the conditions that could render them easy or at least tractable in practice, often other types of approaches have to be attempted.

Although all approaches may be valid and valuable, in many cases one only is explored. More in general, there appears to be relatively little cross-talk and cross-fertilisation being attempted between these different approaches. Guided by problems from computational biology, the goal of this axis is to add to the growing insights on how well such problems can be solved theoretically.

GENSCALE Project-Team

3. Research Program

3.1. Axis 1: Data Structure

The aim of this axis is to develop efficient data structures for representing the mass of genomic data generated by the sequencing machines. This research is motivated by the fact that the treatments of large genomes, such as mammalian or plant genomes, require high computing resources, and more specifically very important memory configuration. For example, the ABYSS software used 4.3TB of memory to assemble the white spruce genome [45]. The main reason for such memory consumption is that the data structures used in ABYSS are far from optimal (and this is also the case for many assembly software).

Our research focuses on the de-Bruijn graph structure. This well-known data structure, directly built from raw sequencing data, have many properties matching perfectly well with NGS processing requirements (see next section). Here, the question we are interested in is how to provide a low memory footprint implementation of the de-Bruijn graph to process very large NGS datasets, including metagenomic ones.

Another research direction of this axis is the indexing of large sets of objects. A typical, but non exclusive, need is to annotate nodes of the de-Bruijn graph, that is potentially billions of items. Again, very low memory footprint indexing structures are mandatory to manage a very large quantity of objects.

3.2. Axis 2: Algorithms

The main goal of the GenScale team is to develop optimized tools dedicated to NGS processing. Optimization can be seen both in terms of space (low memory footprint) and in terms of time (fast execution time). The first point is mainly related to advanced data structures as presented in the previous section (axis 1). The second point relies on new algorithms and, when possible implementation on parallel structures (axis 3).

We do not have the ambition to cover the vast panel of software related to NGS needs. We particularly focused on the following areas:

- **NGS data Compression** De-Bruijn graphs are de facto a compressed representation of the NGS information from which very efficient and specific compressors can be designed. Furthermore, compressing the data using smart structures may speed up some downstream graph-based analyses since a graph structure is already built.
- **Genome assembly** This task remains very complicated, especially for large and complex genomes, such as plant genomes with polyploid and highly repeated structures. We worked both on the generation of contigs and on the scaffolding step.
- **Detection of variants** This is often the first information we want to extract from billions of reads. Variant structures range from SNPs or short indels to large insertions/deletions and long inversions over the chromosomes. We developed original methods to find variants without any reference genome.
- **Metagenomics** We focussed our research on comparative metagenomics by providing methods able to compare hundreds of metagenomic samples together. This is achieved by combining very low memory data structures and efficient implementation and parallelization on large clusters.
- **Genome Wide Association Study (GWAS)** We tackle this problem with algorithms commonly used in data mining. From two cohorts of individuals (case and control) we can exhibit statistically significant *patterns* spanning over full genomes.

3.3. Axis 3: Parallelism

This third axis is another lever to increase performances and scalability of NGS treatments. There are many levels of parallelism that can be used and/or combined to reduce the execution time of very time-consuming bioinformatics processes. A first level is the parallel nature of today processors that now house several cores. A second level is the grid structure that is present in all bioinformatics centers or in the cloud. These two levels are generally combined: a node of a grid is often a multicore system. Another possibility is to add hardware accelerators to a processor. A GPU board is a good example.

GenScale does not do explicit research on parallelism. It exploits the capacity of computing resources to support parallelism. The problem is addressed in two different directions. The first is an engineering approach that uses existing parallel tools to implement algorithms such as multithreading or MapReduce techniques. The second is a parallel algorithmic approach: during the development step, the algorithms are constrained by parallel criteria. This is particularly true for parallel algorithms targeting hardware accelerators.

IBIS Project-Team

3. Research Program

3.1. Analysis of qualitative dynamics of gene regulatory networks

Participants: Hidde de Jong [Correspondent], Michel Page, Delphine Ropers.

The dynamics of gene regulatory networks can be modeled by means of ordinary differential equations (ODEs), describing the rate of synthesis and degradation of the gene products as well as regulatory interactions between gene products and metabolites. In practice, such models are not easy to construct though, as the parameters are often only constrained to within a range spanning several orders of magnitude for most systems of biological interest. Moreover, the models usually consist of a large number of variables, are strongly nonlinear, and include different time-scales, which makes them difficult to handle both mathematically and computationally. This has motivated the interest in qualitative models which, from incomplete knowledge of the system, are able to provide a coarse-grained picture of its dynamics.

A variety of qualitative modeling formalisms have been introduced over the past decades. Boolean or logical models, which describe gene regulatory and signalling networks as discrete-time finite-state transition systems, are probably most widely used. The dynamics of these systems are governed by logical functions representing the regulatory interactions between the genes and other components of the system. IBIS has focused on a related, hybrid formalism that embeds the logical functions describing regulatory interactions into an ODE formalism, giving rise to so-called piecewise-linear differential equations (PLDEs, Figure 2). The use of logical functions allows the qualitative dynamics of the PLDE models to be analyzed, even in high-dimensional systems. In particular, the qualitative dynamics can be represented by means of a so-called state transition graph, where the states correspond to (hyper)rectangular regions in the state space and transitions between states arise from solutions entering one region from another.

First proposed by Leon Glass and Stuart Kauffman in the early seventies, the mathematical analysis of PLDE models has been the subject of active research for more than four decades. IBIS has made contributions on the mathematical level, in collaboration with the BIOCORE and BIPOP project-teams, notably for solving problems induced by discontinuities in the dynamics of the system at the boundaries between regions, where the logical functions may abruptly switch from one discrete value to another, corresponding to the (in)activation of a gene. In addition, many efforts have gone into the development of the computer tool GENETIC NETWORK ANALYZER (GNA) and its applications to the analysis of the qualitative dynamics of a variety of regulatory networks in microorganisms. Some of the methodological work underlying GNA, notably the development of analysis tools based on temporal logics and model checking, which was carried out with the Inria project-teams CONVEX (ex-VASY) and POP-ART, has implications beyond PLDE models as they apply to logical and other qualitative models as well.

3.2. Inference of gene regulatory networks from time-series data

Participants: Eugenio Cinquemani [Correspondent], Johannes Geiselmann, Hidde de Jong, Cyril Dutrieux, Stephan Lacour, Yannick Martin, Michel Page, Corinne Pinel, Delphine Ropers.

Measurements of the transcriptome of a bacterial cell by means of DNA microarrays, RNA sequencing, and other technologies have yielded huge amounts of data on the state of the transcriptional program in different growth conditions and genetic backgrounds, across different time-points in an experiment. The information on the time-varying state of the cell thus obtained has fueled the development of methods for inferring regulatory interactions between genes. In essence, these methods try to explain the observed variation in the activity of one gene in terms of the variation in activity of other genes. A large number of inference methods have been proposed in the literature and have been successful in a variety of applications, although a number of difficult problems remain.

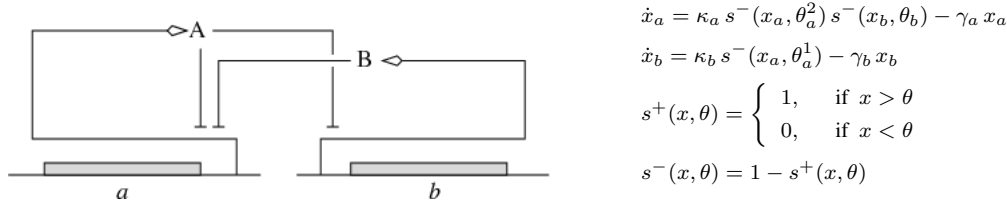


Figure 2. (Left) Example of a gene regulatory network of two genes (a and b), each coding for a regulatory protein (A and B). Protein B inhibits the expression of gene a , while protein A inhibits the expression of gene b and its own gene. (Right) PLDE model corresponding to the network in (a). Protein A is synthesized at a rate κ_a , if and only if the concentration of protein A is below its threshold θ_a^2 ($x_a < \theta_a^2$) and the concentration of protein B below its threshold θ_b ($x_b < \theta_b$). The degradation of protein A occurs at a rate proportional to the concentration of the protein itself ($\gamma_a x_a$).

Current reporter gene technologies, based on Green Fluorescent Proteins (GFPs) and other fluorescent and luminescent reporter proteins, provide an excellent means to measure the activity of a gene *in vivo* and in real time (Figure 3). The underlying principle of the technology is to fuse the promoter region and possibly (part of) the coding region of a gene of interest to a reporter gene. The expression of the reporter gene generates a visible signal (fluorescence or luminescence) that is easy to capture and reflects the expression of a gene of interest. The interest of the reporter systems is further enhanced when they are applied in mutant strains or combined with expression vectors that allow the controlled induction of any particular gene, or the degradation of its product, at a precise moment during the time-course of the experiment. This makes it possible to perturb the network dynamics in a variety of ways, thus obtaining precious information for network inference.

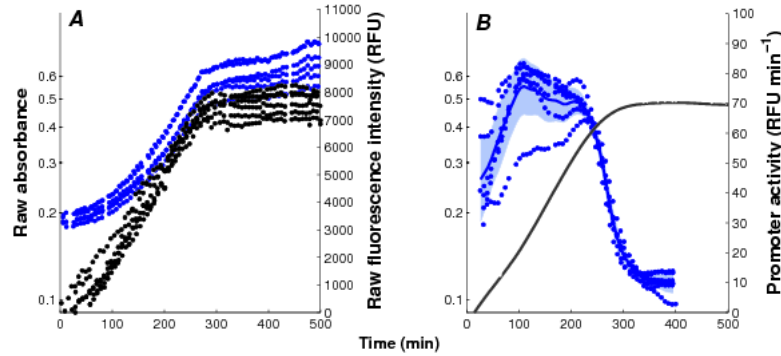


Figure 3. Monitoring of bacterial gene expression *in vivo* using fluorescent reporter genes (Stefan et al., *PLoS Computational Biology*, 11(1):e1004028, 2015). The plots show the primary data obtained in a kinetic experiment with *E. coli* cells, focusing on the expression of the motility gene *tar* in a mutant background. A: Absorbance (●, black) and fluorescence (●, blue) data, corrected for background intensities, obtained with the Δ cpxR strain transformed with the *ptar-gfp* reporter plasmid and grown in M9 with glucose. B: Activity of the *tar* promoter, computed from the primary data. The solid black line corresponds to the mean of 6 replicate absorbance measurements and the shaded blue region to the mean of the promoter activities \pm twice the standard error of the mean.

The specific niche of IBIS in the field of network inference has been the development and application of genome engineering techniques for constructing the reporter and perturbation systems described above, as well as the use of reporter gene data for the reconstruction of gene regulation functions. We have developed an experimental pipeline that resolves most technical difficulties in the generation of reproducible time-series measurements on the population level. The pipeline comes with data analysis software that converts the primary data into measurements of time-varying promoter activities. In addition, for measuring gene expression on the single-cell level by means of microfluidics and time-lapse fluorescence microscopy, we have established collaborations with groups in Grenoble and Paris. The data thus obtained can be exploited for the structural and parametric identification of gene regulatory networks, for which methods with a solid mathematical foundation are developed, in collaboration with colleagues at ETH Zürich and EPF Lausanne (Switzerland). The vertical integration of the network inference process, from the construction of the biological material to the data analysis and inference methods, has the advantage that it allows the experimental design to be precisely tuned to the identification requirements.

3.3. Analysis of integrated metabolic and gene regulatory networks

Participants: Eugenio Cinquemani, Hidde de Jong, Thibault Etienne, Johannes Geiselmann, Stephan Lacour, Yves Markowicz, Michel Page, Corinne Pinel, Delphine Ropers [Correspondent].

The response of bacteria to changes in their environment involves responses on several different levels, from the redistribution of metabolic fluxes and the adjustment of metabolic pools to changes in gene expression. In order to fully understand the mechanisms driving the adaptive response of bacteria, as mentioned above, we need to analyze the interactions between metabolism and gene expression. While often studied in isolation, gene regulatory networks and metabolic networks are closely intertwined. Genes code for enzymes which control metabolic fluxes, while the accumulation or depletion of metabolites may affect the activity of transcription factors and thus the expression of enzyme-encoding genes.

The fundamental principles underlying the interactions between gene expressions and metabolism are far from being understood today. From a biological point of view, the problem is quite challenging, as metabolism and gene expression are dynamic processes evolving on different time-scales and governed by different types of kinetics. Moreover, gene expression and metabolism are measured by different experimental methods generating heterogeneous, and often noisy and incomplete data sets. From a modeling point of view, difficult methodological problems concerned with the reduction and calibration of complex nonlinear models need to be addressed.

Most of the work carried out within the IBIS project-team specifically addressed the analysis of integrated metabolic and gene regulatory networks in the context of *E. coli* carbon metabolism (Figure 4). While an enormous amount of data has accumulated on this model system, the complexity of the regulatory mechanisms and the difficulty to precisely control experimental conditions during growth transitions leave many essential questions open, such as the physiological role and the relative importance of mechanisms on different levels of regulation (transcription factors, metabolic effectors, global physiological parameters, ...). We are interested in the elaboration of novel biological concepts and accompanying mathematical methods to grasp the nature of the interactions between metabolism and gene expression, and thus better understand the overall functioning of the system. Moreover, we have worked on the development of methods for solving what is probably the hardest problem when quantifying the interactions between metabolism and gene expression: the estimation of parameters from heterogeneous and noisy high-throughput data. These problems are tackled in collaboration with experimental groups at Inra/INSA Toulouse and CEA Grenoble, which have complementary experimental competences (proteomics, metabolomics) and biological expertise.

3.4. Natural and engineered control of growth and gene expression

Participants: Célia Boyat, Eugenio Cinquemani, Cyril Dutrieux, Johannes Geiselmann [Correspondent], Hidde de Jong, Stephan Lacour, Ludowic Lancelot, Marco Mauri, Tamas Muszbek, Michel Page, Delphine Ropers.

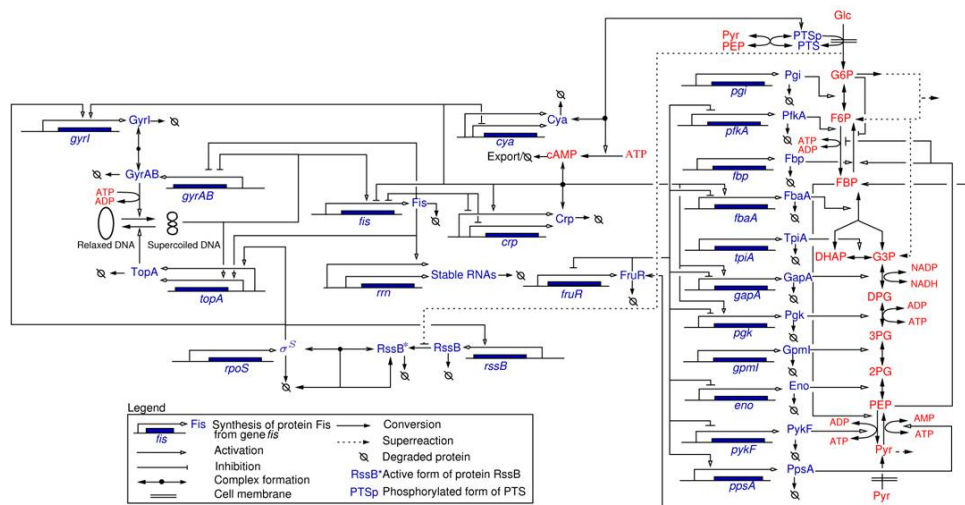


Figure 4. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in *E. coli* (Baldazzi et al., *PLoS Computational Biology*, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the PTS system, via the phosphorylated and non-phosphorylated form of its enzymes (represented by PTSp and PTS, respectively). The pentose-phosphate pathway (PPP) is not explicitly described but we take into account that a small pool of G6P escapes the upper part of glycolysis. At the level of the global regulators the network includes the control of the DNA supercoiling level, the accumulation of the sigma factor RpoS and the Crp-cAMP complex, and the regulatory role exerted by the fructose repressor FruR.

The adaptation of bacterial physiology to changes in the environment, involving changes in the growth rate and a reorganization of gene expression, is fundamentally a resource allocation problem. It notably poses the question how microorganisms redistribute their protein synthesis capacity over different cellular functions when confronted with an environmental challenge. Assuming that resource allocation in microorganisms has been optimized through evolution, for example to allow maximal growth in a variety of environments, this question can be fruitfully formulated as an optimal control problem. We have developed such an optimal control perspective, focusing on the dynamical adaptation of growth and gene expression in response to environmental changes, in close collaboration with the BIOCORE project-team.

A complementary perspective consists in the use of control-theoretical approaches to modify the functioning of a bacterial cell towards a user-defined objective, by rewiring and selectively perturbing its regulatory networks. The question how regulatory networks in microorganisms can be externally controlled using engineering approaches has a long history in biotechnology and is receiving much attention in the emerging field of synthetic biology. Within a number of on-going projects, IBIS is focusing on two different questions. The first concerns the development of open-loop and closed-loop growth-rate controllers of bacterial cells for both fundamental research and biotechnological applications (Figure 5). Second, we are working on the development of methods for the real-time control of gene expression. These methods are obviously capital for the above-mentioned design of growth-rate controllers, but they have also been applied in the context of a platform for real-time control of gene expression in cell population and single cells, developed by the Inria project-team LIFEWARE, in collaboration with a biophysics group at Univ Paris Descartes.

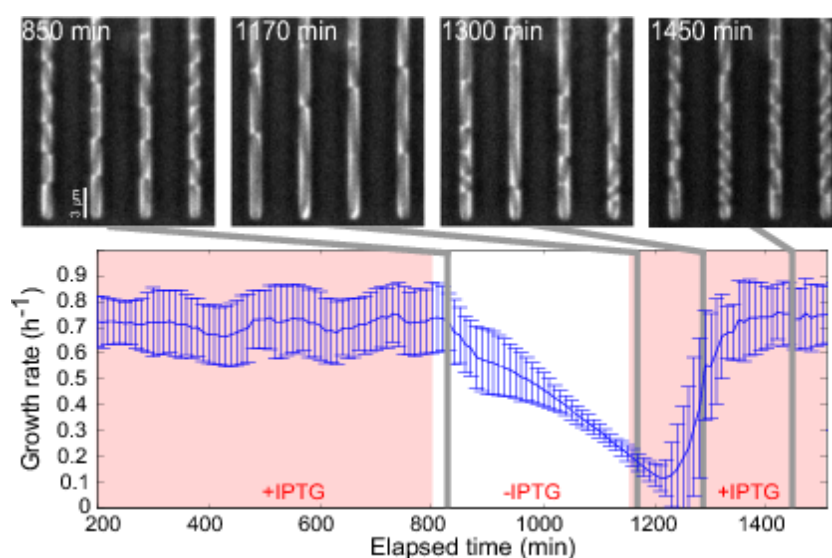


Figure 5. Growth arrest by external control of the gene expression machinery (Izard, Gomez Balderas et al., *Molecular Systems Biology*, 11:840, 2015). An *E. coli* strain in which an essential component of the gene expression machinery, the $\beta\beta'$ subunits of RNA polymerase, was put under the control of an externally-supplied inducer (IPTG), was grown in a microfluidics device and phase-contrast images were acquired every 10 min. The cells were grown in minimal medium with glucose, initially in the presence of 1 mM IPTG. 6 h after removing IPTG from the medium, the growth rate slows down and cells are elongated. About 100 min after adding back 1 mM IPTG into the medium, the elongated cells divide and resume normal growth. The growth rates in the plot are the (weighted) mean of the growth rates of 100 individual cells. The error bars correspond to \pm one standard deviation. The results of the experiment show that the growth rate of a bacterial can be switched off in a reversible manner by an external inducer, based on the reengineering of the natural control of the expression of RNA polymerase.

LIFEWARE Project-Team

3. Research Program

3.1. Computational Systems Biology

Bridging the gap between the complexity of biological systems and our capacity to model and **quantitatively predict system behaviors** is a central challenge in systems biology. We believe that a deeper understanding of the concept and theory of biochemical computation is necessary to tackle that challenge. Progress in the theory is necessary for scaling, and enabling the application of static analysis, module identification and decomposition, model reductions, parameter search, and model inference methods to large biochemical reaction systems. A measure of success on this route will be the production of better computational modeling tools for elucidating the complex dynamics of natural biological processes, designing synthetic biological circuits and biosensors, developing novel therapy strategies, and optimizing patient-tailored therapeutics.

Progress on the **coupling of models to data** is also necessary. Our approach based on quantitative temporal logics provides a powerful framework for formalizing experimental observations and using them as formal specification in model building. Key to success is a tight integration between *in vivo* and *in silico* work, and on the mixing of dry and wet experiments, enabled by novel biotechnologies. In particular, the use of micro-fluidic devices makes it possible to measure behaviors at both single-cell and cell population levels *in vivo*, provided innovative modeling, analysis and control methods are deployed *in silico*.

In synthetic biology, while the construction of simple intracellular circuits has shown feasible, the design of larger, **multicellular systems** is a major open issue. In engineered tissues for example, the behavior results from the subtle interplay between intracellular processes (signal transduction, gene expression) and intercellular processes (contact inhibition, gradient of diffusible molecule), and the question is how should cells be genetically modified such that the desired behavior robustly emerges from cell interactions.

3.2. Chemical Reaction Network (CRN) Theory

Feinberg's chemical reaction network theory and Thomas's influence network analyses provide sufficient and/or necessary structural conditions for the existence of multiple steady states and oscillations in regulatory networks. Those conditions can be verified by static analyzers without knowing kinetic parameter values nor making any simulation. In this domain, most of our work consists in analyzing the interplay between the **structure** (Petri net properties, influence graph, subgraph epimorphisms) and the **dynamics** (Boolean, CTMC, ODE, time scale separations) of biochemical reaction systems. In particular, our study of influence graphs of reaction systems, our generalization of Thomas' conditions of multi-stationarity and Soulé's proof to reaction systems⁰, the inference of reaction systems from ODEs⁰, the computation of structural invariants by constraint programming techniques, and the analysis of model reductions by subgraph epimorphisms now provide solid ground for developing static analyzers, using them on a large scale in systems biology, and elucidating modules.

3.3. Logical Paradigm for Systems Biology

Our group was among the first ones in 2002 to apply **model-checking** methods to systems biology in order to reason on large molecular interaction networks, such as Kohn's map of the mammalian cell cycle (800 reactions over 500 molecules)⁰. The logical paradigm for systems biology that we have subsequently developed for quantitative models can be summarized by the following identifications :

⁰Sylvain Soliman. A stronger necessary condition for the multistationarity of chemical reaction networks. *Bulletin of Mathematical Biology*, 75(11):2289–2303, 2013.

⁰François Fages, Steven Gay, Sylvain Soliman. Inferring reaction systems from ordinary differential equations. *Journal of Theoretical Computer Science (TCS)*, Elsevier, 2015, 599, pp.64–78.

⁰N. Chabrier-Rivier, M. Chiaverini, V. Danos, F. Fages, V. Schächter. Modeling and querying biochemical interaction networks. *Theoretical Computer Science*, 325(1):25–44, 2004.

biological model = transition system K
 dynamical behavior specification = temporal logic formula ϕ
 model validation = model-checking $K, s \models \phi$
 model reduction = sub-model-checking, $K' \subset K$ s.t. $K' \models \phi$
 model prediction = formula enumeration, ϕ s.t. $K, s \models \phi$
 static experiment design = symbolic model-checking, state s s.t. $K, s \models \phi$
 model synthesis = constraint solving $K?, s \models \phi$
 dynamic experiment design = constraint solving $K?, s? \models \phi$

In particular, the definition of a continuous satisfaction degree for **first-order temporal logic** formulae with constraints over the reals, was the key to generalize this approach to quantitative models, opening up the field of model-checking to model optimization⁰ This line of research continues with the development of temporal logic patterns with efficient constraint solvers and their generalization to handle stochastic effects.

3.4. Modeling of Phenotypic Heterogeneity in Cellular Processes

Since nearly two decades, a significant interest has grown for getting a quantitative understanding of the functioning of biological systems at the cellular level. Given their complexity, proposing a model accounting for the observed cell responses, or better, predicting novel behaviors, is now regarded as an essential step to validate a proposed mechanism in systems biology. Moreover, the constant improvement of stimulation and observation tools creates a strong push for the development of methods that provide predictions that are increasingly precise (single cell precision) and robust (complex stimulation profiles).

It is now fully apparent that cells do not respond identically to a same stimulation, even when they are all genetically-identical. This phenotypic heterogeneity plays a significant role in a number of problems ranging from cell resistance to anticancer drug treatments to stress adaptation and bet hedging.

Dedicated modeling frameworks, notably **stochastic** modeling frameworks, such as chemical master equations, and **statistic** modeling frameworks, such as ensemble models, are then needed to capture biological variability.

Appropriate mathematical and computational should then be employed for the analysis of these models and their calibration to experimental data. One can notably mention **global optimization** tools to search for appropriate parameters within large spaces, **moment closure** approaches to efficiently approximate stochastic models⁰, and (stochastic approximations of) the **expectation maximization** algorithm for the identification of mixed-effects models⁰.

3.5. External Control of Cell Processes

External control has been employed since many years to regulate culture growth and other physiological properties. Recently, taking inspiration from developments in synthetic biology, closed loop control has been applied to the regulation of intracellular processes. Such approaches offer unprecedented opportunities to investigate how a cell process dynamical information by maintaining it around specific operating points or driving it out of its standard operating conditions. They can also be used to complement and help the development of synthetic biology through the creation of hybrid systems resulting from the interconnection of in vivo and in silico computing devices.

⁰On a continuous degree of satisfaction of temporal logic formulae with applications to systems biology A. Rizk, G. Batt, F. Fages, S. Soliman International Conference on Computational Methods in Systems Biology, 251-268

⁰Moment-based inference predicts bimodality in transient gene expression, C. Zechner C, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, Proceedings of the National Academy of Sciences USA, 9(5):109(21):8340-5, 2012

⁰What population reveals about individual cell identity: estimation of single-cell models of gene expression in yeast, A. Llamasi, A.M. Gonzalez-Vargas, C. Versari, E. Cinquemani, G. Ferrari-Trecate, P. Hersen, and G. Batt, PLoS Computational Biology, 9(5): e1003056, 2015

In collaboration with Pascal Hersen (CNRS MSC lab), we developed a platform for gene expression control that enables to control protein concentrations in yeast cells. This platform integrates microfluidic devices enabling long-term observation and rapid change of the cells environment, microscopy for single cell measurements, and software for real-time signal quantification and model based control. We demonstrated in 2012 that this platform enables controlling the level of a fluorescent protein in cells with unprecedented accuracy and for many cell generations ⁰.

More recently, motivated by an analogy with a benchmark control problem, the stabilization of an inverted pendulum, we investigated the possibility to balance a genetic toggle switch in the vicinity of its unstable equilibrium configuration. We searched for solutions to balance an individual cell and even an entire population of heterogeneous cells, each harboring a toggle switch [2].

Independently, in collaboration with colleagues from IST Austria, we investigated the problem of controlling cells, one at a time, by constructing an integrated optogenetic platform. It enables experiments that bridge individual and population behaviors. We demonstrated: (i) population structuring by independent closed-loop control of gene expression in many individual cells, (ii) cell-cell variation control during antibiotic perturbation, (iii) hybrid bio-digital circuits in single cells, and freely specifiable digital communication between individual bacteria [1].

3.6. Synthesis of Continuous CRNs

The continuous nature of many protein interactions leads us to consider models of analog computation, and in particular, the recent results in the theory of analog computability and complexity obtained by Amaury Pouly ⁰ and Olivier Bournez, establish fundamental links with digital computation. In [10], we derive from these results a Turing completeness result for elementary reaction systems (without polymerization) under the differential semantics. The proof of this result shows how mathematical functions described by Ordinary Differential Equations, namely by Polynomial Initial Value Problems (PIVP), can be compiled into elementary biochemical reactions, furthermore with a notion of analog computation complexity defined as the length of the trajectory to reach a given precision on the result. This opens a whole research avenue to analyze natural circuits in Systems Biology, transform behavioural specifications into biochemical reactions for Synthetic Biology, and compare artificial circuits with natural circuits acquired through evolution, from the novel point of view of analog computation complexity.

3.7. Constraint Solving and Optimization

Constraint solving and optimization methods are important in our research. On the one hand, static analysis of biochemical reaction networks involves solving hard combinatorial optimization problems, for which **constraint programming** techniques have shown particularly successful, often beating dedicated algorithms and allowing to solve large instances from model repositories. On the other hand, parameter search and model calibration problems involve similarly solving hard continuous optimization problems, for which **evolutionary algorithms** such as the covariance matrix evolution strategy (**CMA-ES**) ⁰ has shown to provide best results in our context, for up to 100 parameters, for building challenging quantitative models, gaining model-based insights, revisiting admitted assumptions, and contributing to biological knowledge ⁰.

⁰Jannis Uhlendorf, Agnès Miermont, Thierry Delaveau, Gilles Charvin, François Fages, Samuel Bottani, Grégory Batt, Pascal Hersen. Long-term model predictive control of gene expression at the population and single-cell levels. *Proceedings of the National Academy of Sciences USA*, 109(35):14271–14276, 2012.

⁰Amaury Pouly, “Continuous models of computation: from computability to complexity”, PhD Thesis, Ecole Polytechnique, Nov. 2015.

⁰N. Hansen, A. Ostermeier (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2) pp. 159–195.

⁰Domitille Heitzler, Guillaume Durand, Nathalie Gallay, Aurélien Rizk, Seungkirl Ahn, Jihee Kim, Jonathan D. Violin, Laurence Dupuy, Christophe Gauthier, Vincent Piketty, Pascale Crépieux, Anne Poupon, Frédérique Clément, François Fages, Robert J. Lefkowitz, Eric Reiter. Competing G protein-coupled receptor kinases balance G protein and β -arrestin signaling. *Molecular Systems Biology*, 8(590), 2012.

MORPHEME Project-Team

3. Research Program

3.1. Research program

The recent advent of an increasing number of new microscopy techniques giving access to high throughput screenings and micro or nano-metric resolutions provides a means for quantitative imaging of biological structures and phenomena. To conduct quantitative biological studies based on these new data, it is necessary to develop non-standard specific tools. This requires using a multi-disciplinary approach. We need biologists to define experiment protocols and interpret the results, but also physicists to model the sensors, computer scientists to develop algorithms and mathematicians to model the resulting information. These different expertises are combined within the Morpheme team. This generates a fecund frame for exchanging expertise, knowledge, leading to an optimal framework for the different tasks (imaging, image analysis, classification, modeling). We thus aim at providing adapted and robust tools required to describe, explain and model fundamental phenomena underlying the morphogenesis of cellular and supra-cellular biological structures. Combining experimental manipulations, in vivo imaging, image processing and computational modeling, we plan to provide methods for the quantitative analysis of the morphological changes that occur during development. This is of key importance as the morphology and topology of mesoscopic structures govern organ and cell function. Alterations in the genetic programs underlying cellular morphogenesis have been linked to a range of pathologies.

Biological questions we will focus on include:

1. what are the parameters and the factors controlling the establishment of ramified structures? (Are they really organize to ensure maximal coverage? How are genetic and physical constraints limiting their morphology?),
2. how are newly generated cells incorporated into reorganizing tissues during development? (is the relative position of cells governed by the lineage they belong to?)

Our goal is to characterize different populations or development conditions based on the shape of cellular and supra-cellular structures, e.g. micro-vascular networks, dendrite/axon networks, tissues from 2D, 2D+t, 3D or 3D+t images (obtained with confocal microscopy, video-microscopy, photon-microscopy or micro-tomography). We plan to extract shapes or quantitative parameters to characterize the morphometric properties of different samples. On the one hand, we will propose numerical and biological models explaining the temporal evolution of the sample, and on the other hand, we will statistically analyze shapes and complex structures to identify relevant markers for classification purposes. This should contribute to a better understanding of the development of normal tissues but also to a characterization at the supra-cellular scale of different pathologies such as Alzheimer, cancer, diabetes, or the Fragile X Syndrome. In this multidisciplinary context, several challenges have to be faced. The expertise of biologists concerning sample generation, as well as optimization of experimental protocols and imaging conditions, is of course crucial. However, the imaging protocols optimized for a qualitative analysis may be sub-optimal for quantitative biology. Second, sample imaging is only a first step, as we need to extract quantitative information. Achieving quantitative imaging remains an open issue in biology, and requires close interactions between biologists, computer scientists and applied mathematicians. On the one hand, experimental and imaging protocols should integrate constraints from the downstream computer-assisted analysis, yielding to a trade-off between qualitative optimized and quantitative optimized protocols. On the other hand, computer analysis should integrate constraints specific to the biological problem, from acquisition to quantitative information extraction. There is therefore a need of specificity for embedding precise biological information for a given task. Besides, a level of generality is also desirable for addressing data from different teams acquired with different protocols and/or sensors. The mathematical modeling of the physics of the acquisition system will yield higher performance reconstruction/restoration algorithms in terms of accuracy. Therefore, physicists and computer scientists have to work together. Quantitative information extraction also has to deal with both the complexity of the structures of interest (e.g., very

dense network, small structure detection in a volume, multiscale behavior, ...) and the unavoidable defects of in vivo imaging (artifacts, missing data, ...). Incorporating biological expertise in model-based segmentation methods provides the required specificity while robustness gained from a methodological analysis increases the generality. Finally, beyond image processing, we aim at quantifying and then statistically analyzing shapes and complex structures (e.g., neuronal or vascular networks), static or in evolution, taking into account variability. In this context, learning methods will be developed for determining (dis)similarity measures between two samples or for determining directly a classification rule using discriminative models, generative models, or hybrid models. Besides, some metrics for comparing, classifying and characterizing objects under study are necessary. We will construct such metrics for biological structures such as neuronal or vascular networks. Attention will be paid to computational cost and scalability of the developed algorithms: biological experiments generally yield huge data sets resulting from high throughput screenings. The research of Morpheme will be developed along the following axes:

- **Imaging:** this includes i) definition of the studied populations (experimental conditions) and preparation of samples, ii) definition of relevant quantitative characteristics and optimized acquisition protocol (staining, imaging, ...) for the specific biological question, and iii) reconstruction/restoration of native data to improve the image readability and interpretation.
- **Feature extraction:** this consists in detecting and delineating the biological structures of interest from images. Embedding biological properties in the algorithms and models is a key issue. Two main challenges are the variability, both in shape and scale, of biological structures and the huge size of data sets. Following features along time will allow to address morphogenesis and structure development.
- **Classification/Interpretation:** considering a database of images containing different populations, we can infer the parameters associated with a given model on each dataset from which the biological structure under study has been extracted. We plan to define classification schemes for characterizing the different populations based either on the model parameters, or on some specific metric between the extracted structures.
- **Modeling:** two aspects will be considered. This first one consists in modeling biological phenomena such as axon growing or network topology in different contexts. One main advantage of our team is the possibility to use the image information for calibrating and/or validating the biological models. Calibration induces parameter inference as a main challenge. The second aspect consists in using a prior based on biological properties for extracting relevant information from images. Here again, combining biology and computer science expertise is a key point.

PLEIADE Team

3. Research Program

3.1. A Geometric View of Diversity

Diversity may be studied as a set of dissimilarities between objects. The underlying mathematical construction is the notion of distance. Knowing a set of objects, it is possible, after computation of pairwise distances, or sometimes dissimilarities, to build a Euclidean image of it as a point cloud in a space of relevant dimension. Then, diversity can be associated with the shape of the point cloud. The human eye is often far better than an algorithm at recognizing a pattern or shape. One objective of our project is to narrow the gap between the story that a human eye can tell, and that an algorithm can tell. Several directions will be explored. First, this requires mastering classical tools in dimension reduction, mainly algebraic tools (PCA, NGS, Isomap, eigenmaps, etc ...). Second, neighborhoods in point clouds naturally lead to graphs describing the neighborhood networks. There is a natural link between modular structures in distance arrays and communities on graphs. Third, points (representing, say, DNA sequences) are samples of diversity. Dimension reduction may show that they live on a given manifold. This leads to geometry (differential or Riemannian geometry). It is expected that some properties of the manifold can tell something of the constraints on the space where measured individuals live. The connection between Riemannian geometry and graphs, where weighted graphs are seen as mesh embedded in a manifold, is currently an active field of research [27], [26]. See as well [29] for a link between geometric structure, linear and nonlinear dimensionality reduction.

Biodiversity and high-performance computing: Most methods and tools for characterizing diversity have been designed for datasets that can be analyzed on a laptop, but NGS datasets produced for metabarcoding are far too large. Data analysis algorithms and tools must be revisited and scaled up. We will mobilize both distributed algorithms like the Arnoldi method and new algorithms, like random projection or column selection methods, to build point clouds in Euclidean spaces from massive data sets, and thus to overcome the cubic complexity of computation of eigenvectors and eigenvalues of very large dense matrices. We will also link distance geometry [23] with convex optimization procedures through matrix completion [19], [20].

Intercalibration: There is a considerable difference between supervised and unsupervised clustering: in supervised clustering, the result for an item i is independent from the result for an item $j \neq i$, whereas in unsupervised clustering, the result for an item i (e.g. the cluster it belongs to, and its composition) depends on nearby items $j \neq i$. Which means that the result may change if some items are added to or subtracted from the sample. This raises the more global problem of how to merge two studies to yield a more comprehensive view of biodiversity?

3.2. Knowledge Management for Biology

The heterogenous data generated in computational molecular biology and ecology are distinguished not only by their volume, but by the richness of the many levels of interpretation that biologists create. The same nucleic acid sequence can be seen as a molecule with a structure, a sequence of base pairs, a collection of genes, an allele, or a molecular fingerprint. To extract the maximum benefit from this treasure trove we must organize the knowledge in ways that facilitate extraction, analysis, and inference. Our focus has been on the efficient representation of relations between biological objects and operations on those representations, in particular heuristic analyses and logical inference.

PLEIADE will develop applications in comparative genomics of related organisms, using new mathematical tools for representing compactly, at different scales of difference, comparisons between related genomes. New methods based on distance geometry will refine these comparisons. Compact representations can be stored, exchanged, and combined. They will form the basis of new simultaneous genome annotation methods, linked directly to abductive inference methods for building functional models of the organisms and their communities.

Since a goal of PLEIADE is to integrate diversity throughout the analysis process, it is necessary to incorporate **diversity as a form of knowledge** that can be stored in a knowledge base. Diversity can be represented using various compact representations, such as trees and quotient graphs storing nested sets of relations. Extracting structured representations and logical relations from integrated knowledge bases (Figure 2) will require domain-specific query methods that can express forms of diversity.

3.3. Modeling by successive refinement

Describing the links between diversity in traits and diversity in function will require comprehensive models, assembled from and refining existing models. A recurring difficulty in building comprehensive models of biological systems is that accurate models for subsystems are built using different formalisms and simulation techniques, and hand-tuned models tend to be so focused in scope that it is difficult to repurpose them [17]. Our belief is that a sustainable effort in building efficient behavioral models must proceed incrementally, rather than by modeling individual processes *de novo*. *Hierarchical modeling* [14] is one way of combining specific models into networks. Effective use of hierarchical models requires both formal definition of the semantics of such composition, and efficient simulation tools for exploring the large space of complex behaviors. We have previously shown that this approach can be effective for certain kinds of systems in biotechnology [2], [18] and medicine [16]. Our challenge is to adapt incremental, hierarchical refinement to modeling organisms and communities in metagenomic and comparative genomic applications.

SERPICO Project-Team

3. Research Program

3.1. Statistics and algorithms for computational microscopy

Fluorescence microscopy limitations are due to the optical aberrations, the resolution of the microscopy system, and the photon budget available for the biological specimen. Hence, new concepts have been defined to address challenging image restoration and molecule detection problems while preserving the integrity of samples. Accordingly, the main stream regarding denoising, deconvolution, registration and detection algorithms advocates appropriate signal processing framework to improve spatial resolution, while at the same time pushing the illumination to extreme low levels in order to limit photo-damages and phototoxicity [7], [6]. As a consequence, the question of adapting cutting-edge signal denoising and deconvolution, object detection, and image registration methods to 3D fluorescence microscopy imaging has retained the attention of several teams over the world.

In this area, the Serpico team has developed a strong expertise in key topics in computational imaging including image denoising and deconvolution, object detection and multimodal image registration. Several algorithms proposed by the team outperformed the state-of-the-art results, and some developments are compatible with “high-throughput microscopy” and the processing of several hundreds of cells. We especially promoted non local, non-parametric and patch-based methods to solve well-known inverse problems or more original reconstruction problems. A recent research direction consists in adapting the deep learning concept to solve challenging detection and reconstruction problems in microscopy. We have investigated convolution neural networks to detect small macromolecules in 3D noisy electron images with promising results. The next step consists in proposing smart paradigms and architectures to save memory and computations.

More generally, many inverse problems and image processing become intractable with modern 3D microscopy, because very large temporal series of volumes (200 to 1000 images per second for one 3D stack) are acquired for several hours. Novel strategies are needed for 3D image denoising, deconvolution and reconstruction since computation is extremely heavy. Accordingly, we will adapt the estimator aggregation approach developed for optical flow computation to meet the requirements of 3D image processing. We plan to investigate regularization-based aggregation energy over super-voxels to reduce complexity, combined to modern optimization algorithms. Finally, we will design parallelized algorithms that fast process 3D images, perform energy minimization in few seconds per image, and run on low-cost graphics processor boards (GPU).

3.2. From image data to motion descriptors: trajectory computation and dynamics analysis

Several particle tracking methods for intracellular analysis have been tailored to cope with different types of cellular and subcellular motion down to Brownian single molecule behavior. Many algorithms were carefully evaluated on the particle tracking challenge dataset published in the Nature Methods journal in 2014 [8]. Actually, there is no definitive solution to the particle tracking problem which remains application-dependent in most cases. The work of Serpico in particle motion analysis is significant in multiple ways, and inserts within a very active international context. One of the remaining key open issues is the tracking of objects with heterogeneous movements in crowded configurations. Moreover, particle tracking methods are not always adapted for motion analysis, especially when the density of moving features hampers the individual extraction of objects of interest undergoing complex motion. Estimating flow fields can be more appropriate to capture the complex dynamics observed in biological sequences. The existing optical flow methods can be classified into two main categories: i/ local methods impose a parametric motion model (e.g. local translation) in a given neighborhood; ii/ global methods estimate the dense motion field by minimizing a global energy functional composed of a data term and a regularization term.

The Serpico team has developed a strong expertise in key topics, especially in object tracking for fluorescence microscopy, optical flow computation and high-level analysis of motion descriptors and trajectories. Several algorithms proposed by the team are very competitive when compared to the state-of-the-art results, and our new paradigms offer promising ways for molecule traffic quantification and analysis. Amongst the problems that we currently address, we can mention: computation of 3D optical flow for large-size images, combination of two frame-based differential methods and sparse sets of trajectories, detection and analysis of unexpected local motion patterns in global coherent collective motion. Development of efficient numerical schemes will be central in the future but visualization methods are also crucial for evaluation and quality assessment. Another direction of research consists in exploiting deep learning to 3D optical flow so as to develop efficient numerical schemes that naturally capture complex motion patterns. Investigation in machine learning and statistics will be actually conducted in the team in the two first research axes to address a large range of inverse problems in bioimaging. Deep learning is an appealing approach since expertise of biologists, via iterative annotation of training data, will be included in the design of image analysis schemes.

3.3. Biological and biophysical models and spatial statistics for quantitative bioimaging

A number of stochastic mathematical models were proposed to describe various intracellular trafficking, where molecules and proteins are transported to their destinations via free diffusion, subdiffusion and ballistic motion representing movements along the cytoskeleton networks assisted by molecular motors. Accordingly, the study of diffusion and stochastic dynamics has known a growing interest in bio-mathematics, biophysics and cell biology with the popularization of fluorescence dynamical microscopy and super-resolution imaging. In this area, the competing teams mainly studied MSD and fluorescence correlation spectroscopy methods.

In the recent period, the Serpico team achieved important results for diffusion-related dynamics involved in exocytosis mechanisms. Robustness to noise has been well investigated, but robustness to environmental effects has yet to be effectively achieved. Particular attention has been given to the estimation of particle motion regime changes, but the available results are still limited for analysing short tracks. The analysis of spatiotemporal molecular interactions from set of 3D computed trajectories or motion vector fields (e.g., co-alignment) must be investigated to fully quantify specific molecular machineries. We have already made efforts in that directions this year (e.g., for colocalization) but important experiments are required to make our preliminary algorithms reliable enough and well adapted to specific transport mechanisms.

Accordingly, we will study quantification methods to represent interactions between molecules and trafficking around three lines of research. First, we will focus on 3D space-time global and local object-based co-orientation and co-alignment methods, in the line of previous work on colocalization, to quantify interactions between molecular species. In addition, given N tracks associated to N molecular species, interaction descriptors, dynamics models and stochastic graphical models representing molecular machines will be studied in the statistical data assimilation framework. Second, we will analyse approaches to estimate molecular mobility, active transport and motion regime changes from computed trajectories in the Lagrangian and Eulerian settings. We will focus on the concept of super-resolution to provide spatially high-resolved maps of diffusion and active transport parameters based on stochastic biophysical models and sparse image representation. Third, we plan to extend the aggregation framework dedicated to optical flow to the problem of diffusion-transport estimation. Finally, we will investigate data assimilation methods to better combine algorithms, models, and experiments in an iterative and virtuous circle. The overview of ultrastructural organization will be achieved by additional 3D electron microscopy technologies.

TAPDANCE Team

3. Research Program

3.1. Recent results and ongoing work

Theoretical work by Meunier and Woods “The non-cooperative tile assembly model is not intrinsically universal or capable of bounded Turing machine simulation” was published in 2017 at the conference STOC, Montreal, Canada, and later presented as a poster and chosen for a short oral talk at DNA23, UT Austin, TX, USA, (presentations by Meunier). In this model, called the noncooperative (or temperature 1) abstract Tile Assembly Model, square tiles assemble into structures in the discrete plane (\mathbb{Z}^2) where each tile binds to a growing structure if one of its 4 coloured edges matches the colour of some available edge on the growing structure. It has been conjectured since 2000 that this model is not capable of computation or other sophisticated forms of growth. We show two results. One of our results states that time-bounded Turing machine computation is impossible in this model if we require the simulation to occur in a bounded rectangle in the plane. This result has a short proof that essentially follows from our other main result (with a more involved proof) which states that this model is not “intrinsically universal”. This latter result means that there is no single tileset in this model that can simulate any instance of the model, answering a question from [Meunier, Patitz, Summers, Theyssier, Winslow, Woods; SODA 2014] and contrasting a result [Doty, Lutz, Patitz, Schweller, Summers, Woods; FOCS 2012] for the more general cooperative (temperature 2) model.

A number of popular models of computation for molecular computing are kinetic in nature: the rules of the systems describe how system evolves, step-by-step over time. However, such models allow one to program structures and behaviours that contradict how molecular systems would behave on long enough (e.g. infinite) time-scales. Recently, Doty, Rogers, Soloveichik, Thachuk and Woods proposed a thermodynamic based model called Thermodynamic Binding Networks. One programs the model by specifying a multiset of molecules, and then the ‘output’ is defined as the (or a, if many) multiset of polymers deemed most ‘thermodynamically stable’ out of all possible multisets of polymers definable in the model. In order to clarify the roles of fundamental thermodynamic concepts in molecular computing, the model makes a number of simplifying assumptions including a lack of geometry and (essentially) an infinite biasing of enthalpy over entropy. This model was published at the DNA23 conference at the University of Texas at Austin in 2017, along with some results: The authors show how to program the model to evaluate Boolean AND/OR formulas, and how to efficiently self-assemble simple structures (a binary counter). Also, limitations (upper bounds) on the size of objects self-assemblable are shown. It is hoped this work will lead to new ways of thinking about computation with molecules.

In experimental work, Thubagere, Li, Johnson, Chen, Doroudi, Lee, Izatt, Wittman, Srinivas, Woods, Winfree and Qian, implemented a molecular walker made out of DNA. The walker randomly around on a 2D nanoscale testing ground, picking up one of two types of cargos (also DNA molecules) and dropping them off at specific goal locations. The work shows the power of simple randomised algorithms in molecular systems. This work was led by Thubagary and Qian at Caltech; Woods’ contribution occurred while at Caltech (before joining Inria). The work was published in Science.

Experimental work by Woods and collaborators Doty, Myhrvold, Hui, Zhou, Yin and Winfree has focused on experimentally implementing a wide class of Boolean circuits of a certain form. Experiments were mostly carried out at Caltech, with data analysis and paper writeup being carried out at Inria, UC Davis and Caltech. A publication is in preparation.

ENS student Tristan Stérin is leading theoretical work on analysing the computational power of the previously mentioned iterated Boolean circuit model of Woods, Doty, Myhrvold, Hui, Zhou, Yin, Winfree. Preliminary results were presented as a poster at DNA23 (UT Austin, TX, USA) where Tristan won “DNA23 best poster award”. The work is in preparation for a conference publication.

There are a number of ongoing projects along the lines of topics above in Overall Objectives.

VIRTUAL PLANTS Project-Team

3. Research Program

3.1. Analysis of structures resulting from meristem activity

To analyze plant growth and structure, we focus mainly on methods for analyzing sequences and tree-structured data. These methods range from algorithms for computing distance between sequences or tree-structured data to statistical models.

- *Combinatorial approaches*: plant structures exhibit complex branching organizations of their organs like internodes, leaves, shoots, axes, branches, etc. These structures can be analyzed with combinatorial methods in order to compare them or to reveal particular types of organization. We investigate a family of techniques to quantify distances between branching systems based on non-linear structural alignment (similar to edit-operation methods used for sequence comparison). Based on these techniques, we study the notion of (topology-based) self-similarity of branching structures in order to define a notion of degree of redundancy for any tree structure and to quantify in this way botanical notions, such as the physiological states of a meristem, fundamental to the description of plant morphogenesis.
- *Statistical modeling*: We investigate different categories of statistical models corresponding to different types of structures.
 - Longitudinal data corresponding to plant growth follow up: the statistical models of interest are equilibrium renewal processes, generalized linear mixed models for longitudinal count data and multiple change-point models for the identification of phenological phases.
 - Repeated patterns within sequences or trees: the statistical models of interest are mainly (hidden) variable-order Markov chains. Hidden variable-order Markov chains were in particular applied to characterize permutation patterns in phyllotaxis and the alternation between flowering and vegetative growth units along sympodial tree axes.
 - Homogeneous zones (or change points) within sequences or trees: most of the statistical models of interest are hidden Markovian models (hidden semi-Markov chains and semi-Markov switching models for sequences and different families of hidden Markov tree models). A complementary approach consists in applying multiple change-point models. The branching structure of a parent shoot is often organized as a succession of branching zones while the succession of shoot at the more macroscopic scale exhibit roughly stationary phases separated by marked change points.

We investigate both estimation methods and diagnostic tools for these different categories of models. In particular we focus on diagnostic tools for latent structure models (e.g. hidden Markovian models or multiple change-point models) that consist in exploring the latent structure space.

- *A new generation of morphogenesis models*: Designing morphogenesis models of the plant development at the macroscopic scales is a challenging problem. As opposed to modeling approaches that attempt to describe plant development on the basis of the integration of purely mechanistic models of various plant functions, we intend to design models that tightly couple mechanistic and empirical sub-models that are elaborated in our plant architecture analysis approach. Empirical models are used as a powerful complementary source of knowledge in places where knowledge about mechanistic processes is lacking or weak. We chose to implement such integrated models in a programming language dedicated to dynamical systems with dynamical structure $(DS)^2$, such as L-systems or MGS. This type of language plays the role of an integration framework for sub-models of heterogeneous nature.

3.2. Meristem functioning and development

In this second scientific axis, we develop models of meristem growth at tissue level in order to integrate various sources of knowledge and to analyze their dynamic and complex spatial interaction. To carry out this integration, we need to develop a complete methodological approach containing:

- algorithms for the automatized segmentation in 3D, and cell lineage tracking throughout time, for images coming from confocal microscopy,
- design of high-level routines and user interfaces to distribute these image analysis tools to the scientific community,
- tools for structural and statistical analysis of 3D meristem structure (spatial statistics, multiscale geometric and topological analysis),
- physical models of cells interactions based on spring-mass systems or on tensorial mechanics at the level of cells,
- models of biochemical networks of hormonal and gene driven regulation, at the cellular and tissue level, using continuous and discrete formalisms,
- and models of cell development taking into account the effects of growth and cell divisions on the two previous classes of models.

ARAMIS Project-Team

3. Research Program

3.1. From geometrical data to multimodal imaging

Brain diseases are associated to alterations of brain structure that can be studied in vivo using anatomical and diffusion MRI. The anatomy of a given subject can be represented by sets of anatomical surfaces (cortical and subcortical surfaces) and curves (white matter tracks) that can be extracted from anatomical and diffusion MRI respectively. We aim to develop approaches that can characterize the variability of brain anatomy within populations of subjects. To that purpose, we propose methods to estimate population atlases that provide an average model of a population of subjects together with a statistical model of their variability. Finally, we aim to introduce representations that can integrate geometrical information (anatomical surfaces, white matter tracts) together with functional (PET, ASL, EEG/MEG) and microstructural information.

3.2. Models of brain networks

Functional imaging techniques (EEG, MEG and fMRI) allow characterizing the statistical interactions between the activities of different brain areas, i.e. functional connectivity. Functional integration of spatially distributed brain regions is a well-known mechanism underlying various cognitive tasks, and is disrupted in brain disorders. Our team develops a framework for the characterization of brain connectivity patterns, based on connectivity descriptors from the theory of complex networks. More specifically, we propose analytical tools to infer brain networks, characterize their structure and integrate multiple networks (for instance from multiple frequency bands or multiple modalities). The genericity of this approach allows us to apply it to various types of data including functional and structural neuroimaging, as well as genomic data.

3.3. Spatiotemporal modeling from longitudinal data

Longitudinal data sets are collected to capture variable temporal phenomena, which may be due to ageing or disease progression for instance. They consist in the observation of several individuals, each of them being observed at multiple points in time. The statistical exploitation of such data sets is notably difficult since data of each individual follow a different trajectory of changes and at its own pace. This difficulty is further increased if observations take the form of structured data like images or measurements distributed at the nodes of a mesh, and if the measurements themselves are normalized data or positive definite matrices for which usual linear operations are not defined. We aim to develop a theoretical and algorithmic framework for learning typical trajectories from longitudinal data sets. This framework is built on tools from Riemannian geometry to describe trajectories of changes for any kind of data and their variability within a group both in terms of the direction of the trajectories and pace.

3.4. Decision support systems

We then aim to develop tools to assist clinical decisions such as diagnosis, prognosis or inclusion in therapeutic trials. To that purpose, we leverage the tools developed by the team, such as multimodal representations, network indices and spatio-temporal models which are combined with advanced classification and regression approaches. We also dedicate strong efforts to rigorous, transparent and reproducible validation of the decision support systems on large clinical datasets.

3.5. Clinical research studies

Finally, we aim to apply advanced computational and statistical tools to clinical research studies. These studies are often performed in collaboration with other researchers of the ICM, clinicians of the Pitié -Salpêtrière hospital or external partners. Notably, our team is very often involved "ex-ante" in clinical research studies. As co-investigators of such studies, we contribute to the definition of objectives, study design and definition of protocols. This is instrumental to perform clinically relevant methodological development and to maximize their medical impact. A large part of these clinical studies were in the field of dementia (Alzheimer's disease, fronto-temporal dementia). Recently, we expanded our scope to other neurodegenerative diseases (Parkinson's disease, multiple sclerosis).

ASCLEPIOS Project-Team

3. Research Program

3.1. Introduction

Tremendous progress has been made in the automated analysis of biomedical images during the past two decades [68]. Readers who are neophytes to the field of medical imaging will find an interesting presentation of acquisition techniques of the main medical imaging modalities in [60], [58]. Regarding target applications, a good review of the state of the art can be found in the book *Computer Integrated Surgery* [56], in N. Ayache's article [63] and in recent review articles [64], [68]. The scientific journals *Medical Image Analysis* [51], *Transactions on Medical Imaging* [57], and *Computer Assisted Surgery* [59] are also good reference material. One can have a good vision of the state of the art from the proceedings of the MICCAI'2010 (Medical Image Computing and Computer Assisted Intervention [54], [55]) and ISBI'2010 (Int. Symp. on Biomedical Imaging [53]) conferences.

For instance, for rigid parts of the body like the head, it is now possible to fuse in a completely automated manner images of the same patient taken from different imaging modalities (e.g. anatomical and functional), or to track the evolution of a pathology through the automated registration and comparison of a series of images taken at distant time instants [69], [79]. It is also possible to obtain from a Magnetic Resonance Image (MRI) of the head a reasonable segmentation of skull tissues, white matter, grey matter, and cerebro-spinal fluid [82], or to measure some functional properties of the heart from dynamic sequences of Magnetic Resonance [62], Ultrasound or Nuclear Medicine images [70].

Despite these advances and successes, statistical models of anatomy are still very crude, resulting in poor registration results in deformable regions of the body, or between different subjects. If some algorithms exploit the physical modeling of the image acquisition process, only a few actually model the physical or even the physiological properties of the human body itself. Coupling biomedical image analysis with anatomical and physiological models of the human body could not only provide a better understanding of observed images and signals, but also more efficient tools for detecting anomalies, predicting evolutions, simulating and assessing therapies.

3.2. Medical Image Analysis

The quality of biomedical images tends to improve constantly (better spatial and temporal resolution, better signal to noise ratio). Not only are the images multidimensional (3 spatial coordinates and possibly one temporal dimension), but medical protocols tend to include multisequence (or multiparametric)⁰ and multimodal images⁰ for each single patient.

⁰Multisequence (or multiparametric) imaging consists in acquiring several images of a given patient with the same imaging modality (e.g. MRI, CT, US, SPECT, etc.) but with varying acquisition parameters. For instance, using MRI, patients followed for multiple sclerosis may undergo every six months a 3D multisequence MR acquisition protocol with different pulse sequences (called T1, T2, PD, Flair, etc.): by varying some parameters of the pulse sequences (e.g Echo Time and Repetition Time), images of the same regions are produced with quite different contrasts depending on the nature and function of the observed structures. In addition, one of the acquisitions (T1) can be combined with the injection of a contrast product (typically Gadolinium) to reveal vessels and some pathologies. Diffusion Tensor Images (DTI) can be acquired to measure the self diffusion of protons in every voxel, allowing the measurement for instance of the direction of white matter fibers in the brain (the same principle can be used to measure the direction of muscular fibers in the heart). Functional MRI of the brain can be acquired by exploiting the so-called Bold Effect (Blood Oxygen Level Dependency): slightly higher blood flow in active regions creates a subtle higher T2* signal which can be detected with sophisticated image processing techniques.

⁰Multimodal acquisition consists in acquiring from the same patient images of different modalities, in order to exploit their complementary nature. For instance, CT and MR may provide information on the anatomy (CT providing contrast between bones and soft tissues while MR within soft tissues of different nature) while SPECT and PET images may provide functional information by measuring a local level of metabolic activity.

Despite remarkable efforts and advances during the past twenty years, the central problems of segmentation and registration have not been solved in the general case. It is our objective in the short term to work on specific versions of these problems, taking into account as much *a priori* information as possible on the underlying anatomy and pathology at hand. It is also our objective to include more knowledge of the physics of image acquisition and observed tissues, as well as of the biological processes involved. Therefore the research activities mentioned in this section will incorporate the advances made in Computational Anatomy and Computational Physiology, as described in sections 3.3 and 3.4.

We plan to pursue our efforts on the following problems:

- multi-dimensional, multi-sequence and multi-modal image segmentation; and
- image Registration/Fusion.

3.3. Computational Anatomy

The aim of Computational Anatomy (CA) is to model and analyse the biological variability of the human anatomy. Typical applications cover the simulation of average anatomies and normal variations, the discovery of structural differences between healthy and diseased populations, and the detection and classification of pathologies from structural anomalies.⁰

Studying the variability of biological shapes is an old problem (cf. the book "On Shape and Growth" by D'Arcy Thompson [81]). Significant efforts have since been made to develop a theory for statistical shape analysis (one can refer to [67] for a good summary, and to the special issue of Neuroimage [80] for recent developments). Despite all these efforts, there are a number of challenging mathematical issues that remain largely unsolved. A particular issue is the computation of statistics on manifolds that can be of infinite dimension (e.g the group of diffeomorphisms).

There is a classical stratification of the problems into the following 3 levels [76]:

1. construction from medical images of anatomical manifolds of points, curves, surfaces and volumes;
2. assignment of a point to point correspondence between these manifolds using a specified class of transformations (e.g. rigid, affine, diffeomorphism);
3. generation of probability laws of anatomical variation from these correspondences.

We plan to focus our efforts on the following problems:

1. statistics on anatomical manifolds;
2. propagation of variability from anatomical manifolds;
3. linking anatomical variability to image analysis algorithms; and
4. grid-computing strategies to exploit large databases.

3.4. Computational Physiology

The objective of Computational Physiology (CP) is to provide models of the major functions of the human body and numerical methods to simulate them. The main applications are in medicine where CP can for instance be used to better understand the basic processes leading to the appearance of a pathology, to model its probable evolution and to plan, simulate, and monitor its therapy.

⁰The NIH has launched in 2005 the Alzheimer's Disease Neuroimaging Initiative (60 million USD), a multi-center MRI study of 800 patients who will be followed during several years. The aim is to establish new surrogate end-points from the automated analysis of temporal sequences, which is a challenging goal for researchers in Computational Anatomy. The data is to be made available to qualified research groups involved or not in the study.

Quite advanced models have already been proposed to study at the molecular, cellular and organ level a number of physiological systems (see for instance [77], [74], [65], [78], [71]). While these models and new ones need to be developed, refined or validated, a grand challenge that we want to address in this project is the automatic adaptation of the model to a given patient by comparing the model with the available biomedical images and signals and possibly also some additional information (e.g. genetic). Building such *patient-specific models* is an ambitious goal, which requires the choice or construction of models with a complexity adapted to the resolution of the accessible measurements and the development of new data assimilation methods coping with massive numbers of measurements and unknowns.

There is a hierarchy of modeling levels for CP models of the human body [66]:

- the first level is mainly geometrical, and addresses the construction of a digital description of the anatomy [61], essentially acquired from medical imagery;
- the second level is physical, involving mainly the biomechanical modeling of various tissues, organs, vessels, muscles and bone structures [72];
- the third level is physiological, involving the modeling of the functions of the major organ systems [73] (e.g. cardiovascular, respiratory, digestive, central or peripheral nervous, muscular, reproductive, hormonal) or some pathological metabolism (e.g. evolution of cancerous or inflammatory lesions, formation of vessel stenoses, etc.); and
- a fourth level is cognitive, modeling the higher functions of the human brain [52].

These different levels of modeling are closely related to each other, and several physiological systems may interact with each other (e.g. the cardiopulmonary interaction [75]). The choice of the resolution at which each level is described is important, and may vary from microscopic to macroscopic, ideally through multiscale descriptions.

Building this complete hierarchy of models is necessary to evolve from a *Visible Human project* (essentially the first level of modeling) to a much more ambitious *Physiological Human project* (see [73], [74]). We will not address all the issues raised by this ambitious project, but instead focus on the topics detailed below. Among them, our objective is to identify some common methods for the resolution of the large inverse problem raised by the coupling of physiological models and medical images for the construction of patient-specific models (e.g. specific variational or sequential methods (EKF), dedicated particle filters). We also plan to develop specific expertise in the extraction of geometrical meshes from medical images for their further use in simulation procedures. Finally, computational models can be used for specific image analysis problems studied in section 3.2 (e.g. segmentation, registration, tracking). Application domains include

1. surgery simulation;
2. cardiac Imaging;
3. brain tumors, neo-angiogenesis, wound healing processes, ovocyte regulation, etc.

3.5. Clinical Validation

If the objective of many of the research activities of the project is the discovery of original methods and algorithms with a proof of its feasibility in a limited number of representative cases (i.e. proofs of concept) and publications in high quality scientific journals, we believe that it is important that a reasonable number of studies include a much more significant validation effort. As the BioMedical Image Analysis discipline becomes more mature, validation is necessary for the transformation of new ideas into clinical tools and/or industrial products. It also helps to get access to larger databases of images and signals, which in turn help to stimulate new ideas and concepts.

ATHENA Project-Team

3. Research Program

3.1. Computational diffusion MRI

Diffusion MRI (dMRI) provides a non-invasive way of estimating in-vivo CNS fiber structures using the average random thermal movement (diffusion) of water molecules as a probe. It's a recent field of research with a history of roughly three decades. It was introduced in the mid 80's by Le Bihan et al [88], Merboldt et al [92] and Taylor et al [102]. As of today, it is the unique non-invasive technique capable of describing the neural connectivity in vivo by quantifying the anisotropic diffusion of water molecules in biological tissues.

3.1.1. Diffusion Tensor Imaging & High Angular Resolution Diffusion Imaging

In dMRI, the acquisition and reconstruction of the diffusion signal allows for the reconstruction of the water molecules displacement probability, known as the Ensemble Average Propagator (EAP) [101], [59]. Historically, the first model in dMRI is the 2nd order diffusion tensor (DTI) [54], [53] which assumes the EAP to be Gaussian centered at the origin. DTI has now proved to be extremely useful to study the normal and pathological human brain [89], [73]. It has led to many applications in clinical diagnosis of neurological diseases and disorder, neurosciences applications in assessing connectivity of different brain regions, and more recently, therapeutic applications, primarily in neurosurgical planning. An important and very successful application of diffusion MRI has been brain ischemia, following the discovery that water diffusion drops immediately after the onset of an ischemic event, when brain cells undergo swelling through cytotoxic edema.

The increasing clinical importance of diffusion imaging has driven our interest to develop new processing tools for Diffusion Tensor MRI. Because of the complexity of the data, this imaging modality raises a large amount of mathematical and computational challenges. We have therefore developed original and efficient algorithms relying on Riemannian geometry, differential geometry, partial differential equations and front propagation techniques to correctly and efficiently estimate, regularize, segment and process Diffusion Tensor MRI (DT-MRI) (see [91] and [90]).

In DTI, the Gaussian assumption over-simplifies the diffusion of water molecules. While it is adequate for voxels in which there is only a single fiber orientation (or none), it breaks for voxels in which there are more complex internal structures and limitates the ability of the DTI to describe complex, singular and intricate fiber configurations (U-shape, kissing or crossing fibers). To overcome this limitation, so-called Diffusion Spectrum Imaging (DSI) [107] and High Angular Resolution Diffusion Imaging (HARDI) methods such as Q-ball imaging [105] and other multi-tensors and compartment models [98], [100], [47], [46], [94] were developed to resolve the orientationnality of more complicated fiber bundle configurations.

Q-Ball imaging (QBI) has been proven very successful in resolving multiple intravoxel fiber orientations in MR images, thanks to its ability to reconstruct the Orientation Distribution Function (ODF, the probability of diffusion in a given direction). These tools play a central role in our work related to the development of a robust and linear spherical harmonic estimation of the HARDI signal and to our development of a regularized, fast and robust analytical QBI solution that outperforms the state-of-the-art ODF numerical technique developed by Tuch. Those contributions are fundamental and have already started to impact on the Diffusion MRI, HARDI and Q-Ball Imaging community [72]. They are at the core of our probabilistic and deterministic tractography algorithms devised to best exploit the full distribution of the fiber ODF (see [69], [3] and [70], [4]).

3.1.2. Beyond DTI with high order tensors

High Order Tensors (HOT) models to estimate the diffusion function while overcoming the shortcomings of the 2nd order tensor model have also been recently proposed such as the Generalized Diffusion Tensor Imaging (G-DTI) model developed by Ozarslan et al [111], [112] or 4th order Tensor Model [52]. For more details, we refer the reader to our articles in [75], [98] where we review HOT models and to our articles

in [90], co-authored with some of our close collaborators, where we review recent mathematical models and computational methods for the processing of Diffusion Magnetic Resonance Images, including state-of-the-art reconstruction of diffusion models, cerebral white matter connectivity analysis, and segmentation techniques. Recently, we started to work on Diffusion Kurtosis Imaging (DKI), of great interest for the company OLEA MEDICAL. Indeed, DKI is fast gaining popularity in the domain for characterizing the diffusion propagator or EAP by its deviation from Gaussianity. Hence it is an important tool in the clinic for characterizing the white-matter's integrity with biomarkers derived from the 3D 4th order kurtosis tensor (KT) [78].

All these powerful techniques are of utmost importance to acquire a better understanding of the CNS mechanisms and have helped to efficiently tackle and solve a number of important and challenging problems [46], [47]. They have also opened up a landscape of extremely exciting research fields for medicine and neuroscience. Hence, due to the complexity of the CNS data and as the magnetic field strength of scanners increase, as the strength and speed of gradients increase and as new acquisition techniques appear [2], these imaging modalities raise a large amount of mathematical and computational challenges at the core of the research we develop at ATHENA [77], [98].

3.1.3. Improving dMRI acquisitions

One of the most important challenges in diffusion imaging is to improve acquisition schemes and analyse approaches to optimally acquire and accurately represent diffusion profiles in a clinically feasible scanning time. Indeed, a very important and open problem in Diffusion MRI is related to the fact that HARDI scans generally require many times more diffusion gradient than traditional diffusion MRI scan times. This comes at the price of longer scans, which can be problematic for children and people with certain diseases. Patients are usually unable to tolerate long scans and excessive motion of the patient during the acquisition process can force a scan to be aborted or produce useless diffusion MRI images. Recently, we have developed novel methods for the acquisition and the processing of diffusion magnetic resonance images, to efficiently provide, with just few measurements, new insights into the structure and anatomy of the brain white matter in vivo.

First, we contributed developing real-time reconstruction algorithm based on the Kalman filter [68]. Then, and more recently, we started to explore the utility of Compressive Sensing methods to enable faster acquisition of dMRI data by reducing the number of measurements, while maintaining a high quality for the results. Compressed Sensing (CS) is a recent technique which has been proved to accurately reconstruct sparse signals from undersampled measurements acquired below the Shannon-Nyquist rate [93].

We have contributed to the reconstruction of the diffusion signal and its important features as the orientation distribution function and the ensemble average propagator, with a special focus on clinical setting in particular for single and multiple Q-shell experiments [93], [60], [62]. Compressive sensing as well as the parametric reconstruction of the diffusion signal in a continuous basis of functions such as the Spherical Polar Fourier basis, have been proved through our recent contributions to be very useful for deriving simple and analytical closed formulae for many important dMRI features, which can be estimated via a reduced number of measurements [93], [60], [62].

We have also contributed to design optimal acquisition schemes for single and multiple q-shell experiments. In particular, the method proposed in [2] helps generate sampling schemes with optimal angular coverage for multi-shell acquisitions. The cost function we proposed is an extension of the electrostatic repulsion to multi-shell and can be used to create acquisition schemes with incremental angular distribution, compatible with prematurely stopped scans. Compared to more commonly used radial sampling, our method improves the angular resolution, as well as fiber crossing discrimination. The optimal sampling schemes, freely available for download⁰, have been selected for use in the HCP (Human Connectome Project)⁰.

We think that such kind of contributions open new perspectives for dMRI applications including, for example, tractography where the improved characterization of the fiber orientations is likely to greatly and quickly help tracking through regions with and/or without crossing fibers [76]

⁰<http://www.emmanuelcaruyer.com/>

⁰<http://humanconnectome.org/documentation/Q1/imaging-protocols.html>

3.1.4. *dMRI modelling, tissue microstructures features recovery & applications*

The dMRI signal is highly complex, hence, the mathematical tools required for processing it have to be commensurate in their complexity. Overall, these last twenty years have seen an explosion of intensive scientific research which has vastly improved and literally changed the face of dMRI. In terms of dMRI models, two trends are clearly visible today: the parametric approaches which attempt to build models of the tissue to explain the signal based on model-parameters such as CHARMED [48], AxCaliber [49] and NODDI [108] to cite but a few, and the non-parametric approaches, which attempt to describe the signal in useful but generic functional bases such as the Spherical Polar Fourier (SPF) basis [51], [50], the Solid Harmonic (SoH) basis [71], the Simple Harmonic Oscillator based Reconstruction and Estimation (SHORE) basis [109] and more recent Mean Apparent Propagator or MAP-MRI basis [110].

However, although great improvements have been made in the last twenty years, major improvements are still required primarily to optimally acquire dMRI data, better understand the biophysics of the signal formation, recover invariant and intrinsic microstructure features, identify bio-physically important bio-markers and improve tractography. For short, there is still considerable room for improvement to take dMRI from the benchside to the bedside.

Therefore, there is still considerable room for improvement when it comes to the concepts and tools able to efficiently acquire, process and analyze the complex structure of dMRI data. Develop ground-breaking tools and models for dMRI is one of the major objectives we would like to achieve in order to lead to a decisive advance and breakthrough in this field.

Then, we propose to investigate the feasibility of using our new models and methods to measure extremely important biological tissue microstructure quantities such as axonal radius and density in white matter. These parameters could indeed provide new insight to better understand the brain's architecture and more importantly could also provide new imaging bio-markers to characterize certain neurodegenerative diseases. This challenging scientific problem, when solved, will lead to direct measurements of important microstructural features that will be integrated in our analysis to provide much greater insight into disease mechanisms, recovery and development. These new microstructural parameters will open the road to go far beyond the limitations of the more simple bio-markers derived from DTI that are clinically used to this date – such as MD and FA which are known to be extremely sensitive to confounding factors such as partial volume and axonal dispersion, non-specific and not able to capture any subtle effects that might be early indicators of diseases [5].

3.1.5. *Towards microstructural based tractography*

In order to go far beyond traditional fiber-tracking techniques, we believe that first order information, i.e. fiber orientations, has to be superseded by second and third order information, such as microstructure details, to improve tractography. However, many of these higher order information methods are relatively new or unexplored and tractography algorithms based on these high order based methods have to be conceived and designed. In this aim, we propose to work with multiple-shells to reconstruct the Ensemble Average Propagator (EAP), which represents the whole 3D diffusion process and use the possibility it offers to deduce valuable insights on the microstructural properties of the white matter. Indeed, from a reconstructed EAP one can compute the angular features of the diffusion in an diffusion Orientation Distribution Function (ODF), providing insight in axon orientation, calculate properties of the entire diffusion in a voxel such as the Mean Squared Diffusivity (MSD) and Return-To-Origin Probability (RTOP), or come forth with bio-markers detailing diffusion along a particular white matter bundle direction such as the Return-to-Axis or Return-to-Plane Probability (RTAP or RTPP). This opens the way to a ground-breaking computational and unified framework for tractography based on EAP and microstructure features [7]. Using additional a priori anatomical [10] and/or functional information, we could also constrain the tractography algorithm to start and terminate the streamlines only at valid processing areas of the brain.

This development of a computational and unified framework for tractography, based on EAP, microstructure and a priori anatomical and/or functional features, will open new perspectives in tractography, paving the way to a new generation of realistic and biologically plausible algorithms able to deal with intricate configurations of white matter fibers and to provide an exquisite and intrinsic brain connectivity quantification.

3.2. MEG and EEG

Electroencephalography (EEG) and Magnetoencephalography (MEG) are two non-invasive techniques for measuring (part of) the electrical activity of the brain. While EEG is an old technique (Hans Berger, a German neuropsychiatrist, measured the first human EEG in 1929), MEG is a rather new one: the first measurements of the magnetic field generated by the electrophysiological activity of the brain were made in 1968 at MIT by D. Cohen. Nowadays, EEG is relatively inexpensive and is routinely used to detect and qualify neural activities (epilepsy detection and characterisation, neural disorder qualification, BCI, ...). MEG is, comparatively, much more expensive as SQUIDS only operate under very challenging conditions (at liquid helium temperature) and as a specially shielded room must be used to separate the signal of interest from the ambient noise. However, as it reveals a complementary vision to that of EEG and as it is less sensitive to the head structure, it also bears great hopes and an increasing number of MEG machines are being installed throughout the world. Inria and ODYSÉE/ATHENA have participated in the acquisition of one such machine installed in the hospital "La Timone" in Marseille.

MEG and EEG can be measured simultaneously (M/EEG) and reveal complementary properties of the electrical fields. The two techniques have temporal resolutions of about the millisecond, which is the typical granularity of the measurable electrical phenomena that arise within the brain. This high temporal resolution makes MEG and EEG attractive for the functional study of the brain. The spatial resolution, on the contrary, is somewhat poor as only a few hundred data points can be acquired simultaneously (about 300-400 for MEG and up to 256 for EEG). MEG and EEG are somewhat complementary with fMRI and SPECT in that those provide a very good spatial resolution but a rather poor temporal resolution (of the order of a second for fMRI and a minute for SPECT). Also, contrarily to fMRI, which "only" measures an haemodynamic response linked to the metabolic demand, MEG and EEG measure a direct consequence of the electrical activity of the brain: it is acknowledged that the signals measured by MEG and EEG correspond to the variations of the post-synaptic potentials of the pyramidal cells in the cortex. Pyramidal neurons compose approximately 80% of the neurons of the cortex, and it requires at least about 50,000 active such neurons to generate some measurable signal.

While the few hundred temporal curves obtained using M/EEG have a clear clinical interest, they only provide partial information on the localisation of the sources of the activity (as the measurements are made on or outside of the head). Thus the practical use of M/EEG data raises various problems that are at the core of the ATHENA research in this topic:

- First, as acquisition is continuous and is run at a rate up to 1kHz, the amount of data generated by each experiment is huge. Data selection and reduction (finding relevant time blocks or frequency bands) and pre-processing (removing artifacts, enhancing the signal to noise ratio, ...) are largely done manually at present. Making a better and more systematic use of the measurements is an important step to optimally exploit the M/EEG data [1].
- With a proper model of the head and of the sources of brain electromagnetic activity, it is possible to simulate the electrical propagation and reconstruct sources that can explain the measured signal. Proposing better models [87], [9] and means to calibrate them [106] so as to have better reconstructions are other important aims of our work.
- Finally, we wish to exploit the temporal resolution of M/EEG and to apply the various methods we have developed to better understand some aspects of the brain functioning, and/or to extract more subtle information out of the measurements. This is of interest not only as a cognitive goal, but it also serves the purpose of validating our algorithms and can lead to the use of such methods in the field of Brain Computer Interfaces. To be able to conduct such kind of experiments, an EEG lab has been set up at ATHENA.

BIOVISION Team

3. Research Program

3.1. Introduction

The Biovision team has started on January 1st, 2016. It aims at developing fundamental research as well as technological developments along two axes.

3.1.1. Axis 1: *High tech vision aid systems for low-vision patients*

Visual impairment affects some 285 million people in the world, mostly in developed countries: 85% have low-vision⁰ or poorer in the better-seeing eye and cannot be corrected or improved with regular eyeglasses (Source: [VisionAware](#)), i.e., have remaining sight. For these people, there is a strong need to conceive new aid-systems to help them in their daily living activities. Such aids already exist and can be divided into two categories according to their function. The first category concerns aids that translate visual information into alternative sensory information, such as touch or sound, called Sensory Substitution Devices (SSDs) [53], [47]. The second category concerns aids that adapt visual information to render it more visible to the patients using scene processing methods and suitable devices. These are based on technological solutions and algorithmic solutions that will enhance important scene characteristics [79], [68]. In Biovision, we focus on this second category by targeting new vision aid systems that will help patients perform the task they primary need in their daily life while investigating solutions which adapt to their own pathology. In our approach, we have these main goals :

1. We aim at developing contacts and collaborations with low-vision center and associations in order to better understand low-vision patients needs and have feedback on our prototypes. We are currently focusing on reading and navigation (indoor or outdoor).
2. We aim at proposing new **scene enhancements** which can be adapted depending on **pathologies**.
3. We want to develop solutions based on head mounted displays and especially low cost and large public systems with full consideration of comfort and **ergonomics**. Our objective is that some of our prototypes could be distributed to patients via transfer or company creation.

3.1.2. Axis 2: *Human vision understanding through joint experimental and modeling studies, for normal and dystrophic retinas*

A holistic point of view is emerging in neuroscience where one can observe simultaneously how vision works at different levels of the hierarchy in the visual system. Multiple scales functional analysis and connectomics are also exploding in brain science, and studies of visual systems are upfront on this fast move. These integrated studies call for new classes of theoretical and integrated models where the goal is the modeling of visual functions such as motion integration.

In Biovision we contribute to a better understanding of the visual system with three main goals:

1. We aim at proposing simplified mathematical models characterizing how the retina converts a visual scene into spike population coding, in normal and under specific pathological conditions.
2. We design biophysical models allowing to better understand the multiscale dynamics of the retina, from dynamics of individual cells to their collective activity, and how changes in biophysical parameters (development, pharmacology, pathology) impacts this dynamics.
3. We want to design an integrated numerical model of the visual stream, with a focus on motion integration, from retina to early visual cortex (V1).
4. We develop a simulation platform emulating the retinal spike-response to visual and prosthetic simulations, in normal and pathological conditions.

⁰low-vision is a condition caused by eye disease, in which visual acuity is 20/70 (meaning that the person is not able to see, at 20 meters from a chart, what a normal person would see at 70 meters).

Finally, although this is not the main goal of our team, another natural avenue of our research will be to develop novel synergistic solutions to solve computer vision tasks based on bio-inspired mechanisms [10].

3.2. Scientific methodology

In this section we briefly describe the scientific methods we use to achieve our research goals.

3.2.1. Adaptive image processing

An impressive range of techniques have been developed in the fields of image processing, computer vision and computer graphics to manipulate and interpret image content for a variety of applications. So far, only a few of these techniques have been applied in the context of vision aid systems and even less have been carefully evaluated with patients. However, it is worth noticing a recent gain of interest from the artificial vision side to low-vision applications⁰. A possible explanation for this comes from the technological side, as new hardware are now available (e.g., head-mounted platforms, depth cams), which make it easier to design prototypes that patients could test.

Enhancing and manipulating image content is the first natural type of image processing to consider. This covers a variety of approaches such as equalization, gamma correction, tone mapping, edge enhancement, image decomposition or cartoonization. Some of these methods have already been tested with low-vision patients [44], [58], [59] or even in retina prosthesis systems as pre-processing steps [43]. More sophisticated approaches have also been tested to help patients focus on the most relevant information, such as scene retargeting [62] and seam carving [48], [45]. Using depth information can be another way to highlight specific features of a scene. Depth can be obtained from stereo head systems or RGB-D cameras. It can be used to make 3D-based enhancements [67], [66] and in the context of assisting visually impaired people, depth information can help them navigating in cluttered environments [64], [52].

Our goal is to investigate which image processing could bring a real advantage to the patients in the design of vision aid systems. We study how to combine them and how to make them adapted to patients pathologies, so that they can not only "see" an image but understand it more efficiently. We work on doing this using virtual and augmented reality technology (Sec. 3.2.2).

3.2.2. Virtual, mixed and augmented reality

Virtual, mixed and augmented reality technology (VR/MR/AR) is becoming an important technology in many areas. It is based on the idea of combining digital words with physical realities in different ways, and it encompasses a wide spectrum of hardware. A new term is increasingly being used, which is cross reality (XR), to refer to this continuum of immersive experiences. Initially pushed by game and film industries, it appears that this technology could change our experiences in many domains such as education, journalism, media, training and healthcare.

To start experiencing XR, some cheap solutions work in conjunction with mobile phones like the **Samsung GR** or the simple **Google Cardboard**. Integrated solutions are also the market such as Oculus, OSVR (open source). New solutions appear on the market with increased capacities. Just to cite a few, let us mention the **Oculus Go**, the **Oculus Rift** or the **HTC vive**. For many of them, it is also planed to have eye tracking inside, as done for example in the **FOVE**.

Given this recent evolution and the promising perspectives, it is our conviction that this technology will play a major role in the domain of low-vision. Not only it can be useful to design novel vision aid systems and rehabilitation programs, but also this technology can help us to study the behaviour of low-vision people more precisely. Following these goals, we develop new immersive experiences to help low-vision people in their daily tasks.

⁰See, e.g., the **Special issue on Assistive Computer Vision and Robotics - "Assistive Solutions for Mobility, Communication and HMI"** from Computer Vision and Image Understanding (August 2016) or the International Workshop on Assistive Computer Vision and Robotics (**ECCV 2016 Satellite workshop**)

3.2.3. *Biophysical modeling*

Modeling in neuroscience has to cope with several competing objectives. On one hand describing the biological realm as close as possible, and, on the other hand, providing tractable equations at least at the descriptive level (simulation, qualitative description) and, when possible, at the mathematical level (i.e., affording a rigorous description). These objectives are rarely achieved simultaneously and most of the time one has to make compromises. In Biovision team we adopt the point of view of physicists: try to capture the phenomenological description of a biophysical mechanism, removing irrelevant details in the description, and try to have a qualitative description of equations behaviour at least at the numerical simulation level, and, when possible, get out analytic results. We do not focus on mathematical proofs, instead insisting on the quality of the model in predicting, and, if possible proposing new experiments. This requires a constant interaction with neuroscientists so as to keep the model on the tracks, warning of too crude approximation, still trying to construct equations from canonical principles [4], [5], [36].

3.2.4. *Methods from theoretical physics*

Biophysical models mainly consist of differential equations (ODEs or PDEs) or integro-differential equations (neural fields). We study them using dynamical systems and bifurcation theory as well as techniques coming from nonlinear physics (amplitude equations, stability analysis, Lyapunov spectrum, correlation analysis, multi-scales methods).

For the study of large scale populations (e.g., when studying population coding) we use methods coming from statistical physics. This branch of physics gave birth to mean-field methods as well statistical methods for large population analysis. We use both of them. Mean-field methods are applied for large scale activity in the retina and in the cortex [7], [11], [46].

For the study of retina population coding we use the so-called Gibbs distribution, initially introduced by Boltzmann and Gibbs. This concept includes, but *is not limited to*, maximum entropy models [63] used by numerous authors in the context of the retina (see, e.g., [71], [73], [61], [60], [74]). These papers were restricted to a statistical description without memory neither causality: the time correlations between successive times is not considered. A paradigmatic example of this is the Ising model, used to describe the retinal activity in, e.g., [71], [73]. However, maximum entropy extends to spatio-temporal correlations as we have shown in, e.g., [5] [75], [33].

More generally, while maximum entropy models rely heavily on the questionable assumption of stationarity, the concept of Gibbs distribution does not need this hypothesis. Besides, it allows to handle models with large memory; it also provides a framework to model anticipation [17]. It includes as well existing models to explain retina statistics such as the Generalized Linear Model (GLM) [51].

CAMIN Team

3. Research Program

3.1. Exploration and understanding of the origins and control of movement

One of CAMIN's areas of expertise is **motion measurement, observation and modeling** in the context of **sensorimotor deficiencies**. The team has the capacity to design advanced protocols to explore motor control mechanisms in more or less invasive conditions in both animal and human.

Human movement can be assessed by several noninvasive means, from motion observation (MOCAP, IMU) to electrophysiological measurements (afferent ENG, EMG, see below). Our general approach is to develop solutions that are realistic in terms of clinical or home use by clinical staff and/or patients for diagnosis and assessment purposes. In doing so, we try to gain a better understanding of motor control mechanisms, including deficient ones, which in turn will give us greater insight into the basics of human motor control. Our ultimate goal is to optimally match a neuroprosthesis to the targeted sensorimotor deficiency.

The team is involved in research projects including:

- **Peripheral nervous system (PNS) exploration, modeling and electrophysiology techniques**
Electroneurography (ENG) and electromyography (EMG) signals inform about neural and muscular activities. The team investigates both natural and evoked ENG/EMG through advanced and dedicated signal processing methods. Evoked responses to ES are very precious information for understanding neurophysiological mechanisms, as both the input (ES) and the output (evoked EMG/ENG) are controlled. CAMIN has the expertise to perform animal experiments (rabbits, rats, earthworms and big animals with partners), design hardware and software setups to stimulate and record in harsh conditions, process signals, analyze results and develop models of the observed mechanisms. Experimental surgery is mandatory in our research prior to invasive interventions in humans. It allows us to validate our protocols from theoretical, practical and technical aspects.
- **Central nervous system (CNS) exploration**
Stimulating the CNS directly instead of nerves allows activation of the neural networks responsible for generating functions. Once again, if selectivity is achieved the number of implanted electrodes and cables would be reduced, as would the energy demand. We have investigated **spinal electrical stimulation** in animals (pigs) for urinary track and lower limb function management. This work is very important in terms of both future applications and the increase in knowledge about spinal circuitry. The challenges are technical, experimental and theoretical, and the preliminary results have enabled us to test some selectivity modalities through matrix electrode stimulation. This research area will be further intensified in the future as one of ways to improve neuroprosthetic solutions. We intend to gain a better understanding of the electrophysiological effects of DES through electroencephalographic (EEG) and electrocorticographic (ECoG) recordings in order to optimize anatomo-functional brain mapping, better understand brain dynamics and plasticity, and improve surgical planning, rehabilitation, and the quality of life of patients.
- **Muscle models and fatigue exploration**
Muscle fatigue is one of the major limitations in all FES studies. Simply, the muscle torque varies over time even when the same stimulation pattern is applied. As there is also muscle recovery when there is a rest between stimulations, modeling the fatigue is almost an impossible task. Therefore, it is essential to monitor the muscle state and assess the expected muscle response by FES to improve the current FES system in the direction of greater adaptive force/torque control in the presence of muscle fatigue.
- **Movement interpretation**

We intend to develop ambulatory solutions to allow ecological observation. We have extensively investigated the possibility of using inertial measurement units (IMUs) within body area networks to observe movement and assess posture and gait variables. We have also proposed extracting gait parameters like stride length and foot-ground clearance for evaluation and diagnosis purposes.

3.2. Movement assistance and/or restoration

The challenges in movement restoration are: (i) improving nerve/muscle stimulation modalities and efficiency and (ii) global management of the function that is being restored in interaction with the rest of the body under voluntary control. For this, both local (muscle) and global (function) controls have to be considered.

Online modulation of ES parameters in the context of lower limb functional assistance requires the availability of information about the ongoing movement. Different levels of complexity can be considered, going from simple open-loop to complex control laws (figure 2).

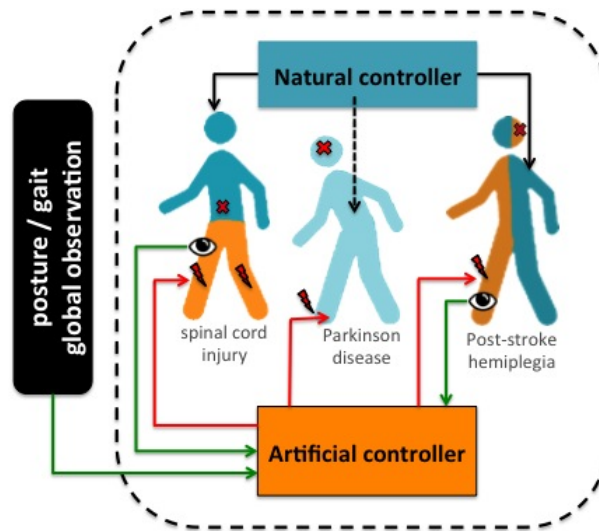


Figure 2. FES assistance should take into account the coexistence of artificial and natural controllers. Artificial controllers should integrate both global (posture/gait) and local (limb/joint) observations.

Real-time adaptation of the stimulation patterns is an important challenge in most of the clinical applications we consider. The modulation of ES parameters to adapt to the occurrence of muscular fatigue or to environment changes needs for advanced adaptive controllers based on sensory information. A special care in minimizing the number of sensors and their impact on patient motion should be taken.

GALEN Project-Team

3. Research Program

3.1. Shape, Grouping and Recognition

A general framework for the fundamental problems of image segmentation, object recognition and scene analysis is the interpretation of an image in terms of a set of symbols and relations among them. Abstractly stated, image interpretation amounts to mapping an observed image, X to a set of symbols Y . Of particular interest are the symbols Y^* that *optimally explain the underlying image*, as measured by a scoring function s that aims at distinguishing correct (consistent with human labellings) from incorrect interpretations:

$$Y^* = \operatorname{argmax}_Y s(X, Y) \quad (4)$$

Applying this framework requires (a) identifying which symbols and relations to use (b) learning a scoring function s from training data and (c) optimizing over Y in Eq.1 .

One of the main themes of our work is the development of methods that jointly address (a,b,c) in a shape-grouping framework in order to reliably extract, describe, model and detect shape information from natural and medical images. A principal motivation for using a shape-based framework is the understanding that shape- and more generally, grouping- based representations can go all the way from image features to objects. Regarding aspect (a), image representation, we cater for the extraction of image features that respect the shape properties of image structures. Such features are typically constructed to be purely geometric (e.g. boundaries, symmetry axes, image segments), or appearance-based, such as image descriptors. The use of machine learning has been shown to facilitate the robust and efficient extraction of such features, while the grouping of local evidence is known to be necessary to disambiguate the potentially noisy local measurements. In our research we have worked on improving feature extraction, proposing novel blends of invariant geometric- and appearance- based features, as well as grouping algorithms that allow for the efficient construction of optimal assemblies of local features.

Regarding aspect (b) we have worked on learning scoring functions for detection with deformable models that can exploit the developed low-level representations, while also being amenable to efficient optimization. Our works in this direction build on the graph-based framework to construct models that reflect the shape properties of the structure being modeled. We have used discriminative learning to exploit boundary- and symmetry-based representations for the construction of hierarchical models for shape detection, while for medical images we have developed methods for the end-to-end discriminative training of deformable contour models that combine low-level descriptors with contour-based organ boundary representations.

Regarding aspect (c) we have developed algorithms which implement top-down/bottom-up computation both in deterministic and stochastic optimization. The main idea is that ‘bottom-up’, image-based guidance is necessary for efficient detection, while ‘top-down’, object-based knowledge can disambiguate and help reliably interpret a given image; a combination of both modes of operation is necessary to combine accuracy with efficiency. In particular we have developed novel techniques for object detection that employ combinatorial optimization tools (A* and Branch-and-Bound) to tame the combinatorial complexity, achieving a best-case performance that is logarithmic in the number of pixels.

In the long run we aim at scaling up shape-based methods to 3D detection and pose estimation and large-scale object detection. One aspect which seems central to this is the development of appropriate mid-level representations. This is a problem that has received increased interest lately in the 2D case and is relatively mature, but in 3D it has been pursued primarily through ad-hoc schemes. We anticipate that questions pertaining to part sharing in 3D will be addressed most successfully by relying on explicit 3D representations. On the one hand depth sensors, such as Microsoft's Kinect, are now cheap enough to bring surface modeling and matching into the mainstream of computer vision - so these advances may be directly exploitable at test time for detection. On the other hand, even if we do not use depth information at test time, having 3D information can simplify the modeling task during training. In on-going work with collaborators we have started exploring combinations of such aspects, namely (i) the use of surface analysis tools to match surfaces from depth sensors (ii) using branch-and-bound for efficient inference in 3D space and (iii) groupwise-registration to build statistical 3D surface models. In the coming years we intend to pursue a tighter integration of these different directions for scalable 3D object recognition.

3.2. Machine Learning & Structured Prediction

The foundation of statistical inference is to learn a function that minimizes the expected loss of a prediction with respect to some unknown distribution

$$\mathcal{R}(f) = \int \ell(f, x, y) dP(x, y), \quad (5)$$

where $\ell(f, x, y)$ is a problem specific loss function that encodes a penalty for predicting $f(x)$ when the correct prediction is y . In our case, we consider x to be a medical image, and y to be some prediction, e.g. the segmentation of a tumor, or a kinematic model of the skeleton. The loss function, ℓ , is informed by the costs associated with making a specific misprediction. As a concrete example, if the true spatial extent of a tumor is encoded in y , $f(x)$ may make mistakes in classifying healthy tissue as a tumor, and mistakes in classifying diseased tissue as healthy. The loss function should encode the potential physiological damage resulting from erroneously targeting healthy tissue for irradiation, as well as the risk from missing a portion of the tumor.

A key problem is that the distribution P is unknown, and any algorithm that is to estimate f from labeled training examples must additionally make an implicit estimate of P . A central technology of empirical inference is to approximate $\mathcal{R}(f)$ with the empirical risk,

$$\mathcal{R}(f) \approx \widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, x_i, y_i), \quad (6)$$

which makes an implicit assumption that the training samples (x_i, y_i) are drawn i.i.d. from P . Direct minimization of $\widehat{\mathcal{R}}(f)$ leads to overfitting when the function class $f \in \mathcal{F}$ is too rich, and regularization is required:

$$\min_{f \in \mathcal{F}} \lambda \Omega(\|f\|) + \widehat{\mathcal{R}}(f), \quad (7)$$

where Ω is a monotonically increasing function that penalizes complex functions.

Equation Eq. 4 is very well studied in classical statistics for the case that the output, $y \in \mathcal{Y}$, is a binary or scalar prediction, but this is not the case in most medical imaging prediction tasks of interest. Instead, complex interdependencies in the output space leads to difficulties in modeling inference as a binary prediction problem. One may attempt to model e.g. tumor segmentation as a series of binary predictions at each voxel in a medical image, but this violates the i.i.d. sampling assumption implicit in Equation Eq. 3. Furthermore, we typically gain performance by appropriately modeling the inter-relationships between voxel predictions, e.g. by incorporating pairwise and higher order potentials that encode prior knowledge about the problem domain. It is in this context that we develop statistical methods appropriate to structured prediction in the medical imaging setting.

3.3. Self-Paced Learning with Missing Information

Many tasks in artificial intelligence are solved by building a model whose parameters encode the prior domain knowledge and the likelihood of the observed data. In order to use such models in practice, we need to estimate its parameters automatically using training data. The most prevalent paradigm of parameter estimation is supervised learning, which requires the collection of the inputs x_i and the desired outputs y_i . However, such an approach has two main disadvantages. First, obtaining the ground-truth annotation of high-level applications, such as a tight bounding box around all the objects present in an image, is often expensive. This prohibits the use of a large training dataset, which is essential for learning the existing complex models. Second, in many applications, particularly in the field of medical image analysis, obtaining the ground-truth annotation may not be feasible. For example, even the experts may disagree on the correct segmentation of a microscopical image due to the similarities between the appearance of the foreground and background.

In order to address the deficiencies of supervised learning, researchers have started to focus on the problem of parameter estimation with data that contains hidden variables. The hidden variables model the missing information in the annotations. Obtaining such data is practically more feasible: image-level labels ('contains car', 'does not contain person') instead of tight bounding boxes; partial segmentation of medical images. Formally, the parameters \mathbf{w} of the model are learned by minimizing the following objective:

$$\min_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w}) + \sum_{i=1}^n \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \quad (8)$$

Here, \mathcal{W} represents the space of all parameters, n is the number of training samples, $R(\cdot)$ is a regularization function, and $\Delta(\cdot)$ is a measure of the difference between the ground-truth output y_i and the predicted output and hidden variable pair $(y_i(\mathbf{w}), h_i(\mathbf{w}))$.

Previous attempts at minimizing the above objective function treat all the training samples equally. This is in stark contrast to how a child learns: first focus on easy samples ('learn to add two natural numbers') before moving on to more complex samples ('learn to add two complex numbers'). In our work, we capture this intuition using a novel, iterative algorithm called self-paced learning (SPL). At an iteration t , SPL minimizes the following objective function:

$$\min_{\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \{0,1\}^n} R(\mathbf{w}) + \sum_{i=1}^n v_i \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})) - \mu_t \sum_{i=1}^n v_i. \quad (9)$$

Here, samples with $v_i = 0$ are discarded during the iteration t , since the corresponding loss is multiplied by 0. The term μ_t is a threshold that governs how many samples are discarded. It is annealed at each iteration, allowing the learner to estimate the parameters using more and more samples, until all samples are used. Our results already demonstrate that SPL estimates accurate parameters for various applications such as image classification, discriminative motif finding, handwritten digit recognition and semantic segmentation. We will investigate the use of SPL to estimate the parameters of the models of medical imaging applications, such as segmentation and registration, that are being developed in the GALEN team. The ability to handle missing information is extremely important in this domain due to the similarities between foreground and background appearances (which results in ambiguities in annotations). We will also develop methods that are capable of minimizing more general loss functions that depend on the (unknown) value of the hidden variables, that is,

$$\min_{\mathbf{w} \in \mathcal{W}, \theta \in \Theta} R(\mathbf{w}) + \sum_{i=1}^n \sum_{h_i \in \mathcal{H}} \Pr(h_i | x_i, y_i; \theta) \Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \quad (10)$$

Here, θ is the parameter vector of the distribution of the hidden variables h_i given the input x_i and output y_i , and needs to be estimated together with the model parameters \mathbf{w} . The use of a more general loss function will allow us to better exploit the freely available data with missing information. For example, consider the case where y_i is a binary indicator for the presence of a type of cell in a microscopical image, and h_i is a tight bounding box around the cell. While the loss function $\Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w}))$ can be used to learn to classify an image as containing a particular cell or not, the more general loss function $\Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w}))$ can be used to learn to detect the cell as well (since h_i models its location)

3.4. Discrete Biomedical Image Perception

A wide variety of tasks in medical image analysis can be formulated as discrete labeling problems. In very simple terms, a discrete optimization problem can be stated as follows: we are given a discrete set of variables \mathcal{V} , all of which are vertices in a graph \mathcal{G} . The edges of this graph (denoted by \mathcal{E}) encode the variables' relationships. We are also given as input a discrete set of labels \mathcal{L} . We must then assign one label from \mathcal{L} to each variable in \mathcal{V} . However, each time we choose to assign a label, say, x_{p_1} to a variable p_1 , we are forced to pay a price according to the so-called *singleton* potential function $g_p(x_p)$, while each time we choose to assign a pair of labels, say, x_{p_1} and x_{p_2} to two interrelated variables p_1 and p_2 (two nodes that are connected by an edge in the graph \mathcal{G}), we are also forced to pay another price, which is now determined by the so called *pairwise* potential function $f_{p_1 p_2}(x_{p_1}, x_{p_2})$. Both the singleton and pairwise potential functions are problem specific and are thus assumed to be provided as input.

Our goal is then to choose a labeling which will allow us to pay the smallest total price. In other words, based on what we have mentioned above, we want to choose a labeling that minimizes the sum of all the MRF potentials, or equivalently the MRF energy. This amounts to solving the following optimization problem:

$$\arg \min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}). \quad (11)$$

The use of such a model can describe a number of challenging problems in medical image analysis. However these simplistic models can only account for simple interactions between variables, a rather constrained scenario for high-level medical imaging perception tasks. One can augment the expression power of this model through higher order interactions between variables, or a number of cliques $\{C_i, i \in [1, n] = \{\{p_{i1}, \dots, p_{i|C_i|}\}\}$ of order $|C_i|$ that will augment the definition of \mathcal{V} and will introduce hyper-vertices:

$$\arg \min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}) + \sum_{C_i \in \mathcal{C}} f_{p_1 \dots p_n}(x_{p_{i1}}, \dots, x_{p_{i|C_i|}}). \quad (12)$$

where $f_{p_1 \dots p_n}$ is the price to pay for associating the labels $(x_{p_{i1}}, \dots, x_{p_{i|C_i|}})$ to the nodes $(p_1 \dots p_{i|C_i|})$. Parameter inference, addressed by minimizing the problem above, is the most critical aspect in computational medicine and efficient optimization algorithms are to be evaluated both in terms of computational complexity as well as of inference performance. State of the art methods include deterministic and non-deterministic annealing, genetic algorithms, max-flow/min-cut techniques and relaxation. These methods offer certain strengths while exhibiting certain limitations, mostly related to the amount of interactions which can be tolerated among neighborhood nodes. In the area of medical imaging where domain knowledge is quite strong, one would expect that such interactions should be enforced at the largest scale possible.

MATHNEURO Team

3. Research Program

3.1. Neural networks dynamics

The study of neural networks is certainly motivated by the long term goal to understand how brain is working. But, beyond the comprehension of brain or even of simpler neural systems in less evolved animals, there is also the desire to exhibit general mechanisms or principles at work in the nervous system. One possible strategy is to propose mathematical models of neural activity, at different space and time scales, depending on the type of phenomena under consideration. However, beyond the mere proposal of new models, which can rapidly result in a plethora, there is also a need to understand some fundamental keys ruling the behaviour of neural networks, and, from this, to extract new ideas that can be tested in real experiments. Therefore, there is a need to make a thorough analysis of these models. An efficient approach, developed in our team, consists of analysing neural networks as dynamical systems. This allows to address several issues. A first, natural issue is to ask about the (generic) dynamics exhibited by the system when control parameters vary. This naturally leads to analyse the bifurcations [52] [53] occurring in the network and which phenomenological parameters control these bifurcations. Another issue concerns the interplay between neuron dynamics and synaptic network structure.

3.2. Mean-field and stochastic approaches

Modeling neural activity at scales integrating the effect of thousands of neurons is of central importance for several reasons. First, most imaging techniques are not able to measure individual neuron activity (microscopic scale), but are instead measuring mesoscopic effects resulting from the activity of several hundreds to several hundreds of thousands of neurons. Second, anatomical data recorded in the cortex reveal the existence of structures, such as the cortical columns, with a diameter of about $50\mu m$ to $1mm$, containing of the order of one hundred to one hundred thousand neurons belonging to a few different species. The description of this collective dynamics requires models which are different from individual neurons models. In particular, when the number of neurons is large enough averaging effects appear, and the collective dynamics is well described by an effective mean-field, summarizing the effect of the interactions of a neuron with the other neurons, and depending on a few effective control parameters. This vision, inherited from statistical physics requires that the space scale be large enough to include a large number of microscopic components (here neurons) and small enough so that the region considered is homogeneous.

Our group is developing mathematical and numerical methods allowing on one hand to produce dynamic mean-field equations from the physiological characteristics of neural structure (neurons type, synapse type and anatomical connectivity between neurons populations), and on the other so simulate these equations. These methods use tools from advanced probability theory such as the theory of Large Deviations [45] and the study of interacting diffusions [35].

3.3. Neural fields

Neural fields are a phenomenological way of describing the activity of population of neurons by delayed integro-differential equations. This continuous approximation turns out to be very useful to model large brain areas such as those involved in visual perception. The mathematical properties of these equations and their solutions are still imperfectly known, in particular in the presence of delays, different time scales and noise.

Our group is developing mathematical and numerical methods for analysing these equations. These methods are based upon techniques from mathematical functional analysis, bifurcation theory [54], equivariant bifurcation analysis, delay equations, and stochastic partial differential equations. We have been able to characterize the solutions of these neural fields equations and their bifurcations, apply and expand the theory to account for such perceptual phenomena as edge, texture [37], and motion perception. We have also developed a theory of the delayed neural fields equations, in particular in the case of constant delays and propagation delays that must be taken into account when attempting to model large size cortical areas [55]. This theory is based on center manifold and normal forms ideas.

3.4. Slow-Fast Dynamics in Neuronal Models

Neuronal rhythms typically display many different timescales, therefore it is important to incorporate this slow-fast aspect in models. We are interested in this modeling paradigm where slow-fast point models (using Ordinary Differential Equations) are investigated in terms of their bifurcation structure and the patterns of oscillatory solutions that they can produce. To insight into the dynamics of such systems, we use a mix of theoretical techniques — such as geometric desingularisation and centre manifold reduction [50] — and numerical methods such as pseudo-arclength continuation [43]. We are interested in families of complex oscillations generated by both mathematical and biophysical models of neurons. In particular, so-called *mixed-mode oscillations (MMOs)* [41], [49], which represent an alternation between subthreshold and spiking behaviour, and *bursting oscillations* [42], [47], also corresponding to experimentally observed behaviour [39].

3.5. Modeling neuronal excitability

Excitability refers to the all-or-none property of neurons [44], [48]. That is, the ability to respond nonlinearly to an input with a dramatic change of response from “none” — no response except a small perturbation that returns to equilibrium — to “all” — large response with the generation of an action potential or spike before the neuron returns to equilibrium. The return to equilibrium may also be an oscillatory motion of small amplitude; in this case, one speaks of resonator neurons as opposed to integrator neurons. The combination of a spike followed by subthreshold oscillations is then often referred to as mixed-mode oscillations (MMOs) [41]. Slow-fast ODE models of dimension at least three are well capable of reproducing such complex neural oscillations. Part of our research expertise is to analyse the possible transitions between different complex oscillatory patterns of this sort upon input change and, in mathematical terms, this corresponds to understanding the bifurcation structure of the model. Furthermore, the shape of time series of this sort with a given oscillatory pattern can be analysed within the mathematical framework of dynamic bifurcations; see the section on slow-fast dynamics in Neuronal Models. The main example of abnormal neuronal excitability is hyperexcitability and it is important to understand the biological factors which lead to such excess of excitability and to identify (both in detailed biophysical models and reduced phenomenological ones) the mathematical structures leading to these anomalies. Hyperexcitability is one important trigger for pathological brain states related to various diseases such as chronic migraine, epilepsy or even Alzheimer’s Disease. A central central axis of research within our group is to revisit models of such pathological scenarios, in relation with a combination of advanced mathematical tools and in partnership with biological labs.

3.6. Synaptic Plasticity

Neural networks show amazing abilities to evolve and adapt, and to store and process information. These capabilities are mainly conditioned by plasticity mechanisms, and especially synaptic plasticity, inducing a mutual coupling between network structure and neuron dynamics. Synaptic plasticity occurs at many levels of organization and time scales in the nervous system [36]. It is of course involved in memory and learning mechanisms, but it also alters excitability of brain areas and regulates behavioral states (e.g., transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge.

Our group is developing mathematical and numerical methods to analyse this mutual interaction. On the one hand, we have shown that plasticity mechanisms, Hebbian-like or STDP, have strong effects on neuron dynamics complexity, such as dynamics complexity reduction, and spike statistics

MIMESIS Team

3. Research Program

3.1. Real Time Patient-Specific Computational Models

Accounting for the biomechanics and physiology of organs under various stimuli requires employment of *biomechanical models*. These must describe biophysical phenomena such as soft-tissue deformation, fluid dynamics, electrical propagation, or heat transfer. We aim at simulating the impact of certain therapies (such as cryosurgery or radio-frequency ablation) and representing the behavior of complex organs such as the brain, the liver or the heart.

An important part of our research is dedicated to the development of new accurate models that remain compatible with real-time computation [6]. Such advanced models do not only permit to increase the realism of future training systems, but they act as a bridge toward the development of patient-specific preoperative planning as well as augmented reality tools for the operating room [5] [40], [54]. Yet, patient-specific planning or per-operative guidance also requires the models to be parametrized with patient-specific biomechanical data. Our objective is the study of hyper-elastic models and their validation for a range of tissues. Preliminary work has been done through two collaborations, one with the biomechanical lab in Lille (LML) [42], and the biomechanics group from the Icube laboratory in Strasbourg on the development and validation of liver and kidney models [52].

Another important research topic is related to model reduction through various approaches, such as Proper Generalized Decomposition (PGD) [35]. Similar approaches, such as the use of Krylov spaces, have already been studied in our group recently [33].

We continue our work on cardiac electro-physiology simulation [53], with a focus on patient-specific adaptation of the model. We also study a similar problem, related to the modeling of the electrical conduction in the brain, in the context of Deep Brain Stimulation (DBS) [34], [38]. In this neurosurgical procedure, electrodes are implanted deep into the brain and, connected to a brain pacemaker, send electrical impulses to specific regions. A final objective is to solve optimization problems in the context of heat diffusion. This is a key element of the development of a planning system that can estimate the locations of the electrodes leading to an optimal therapeutic effect [54].

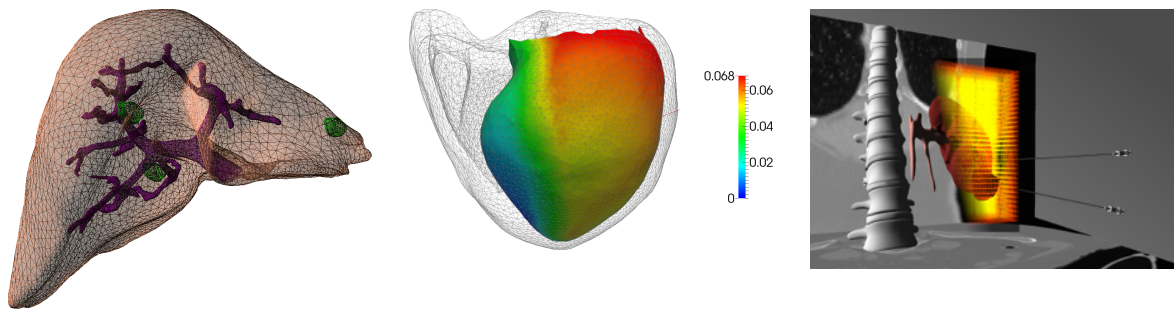


Figure 2. Left: patient-specific liver model with its vascular system. Middle: patient specific depolarization times. Right: cryoablation in the kidney.

3.2. Adaptive Meshing and Advanced Simulation Techniques

Most simulations in the field of biomechanics, physiological modeling, or even computer graphics, are performed using finite element approaches. Such simulations require a discretization of the domain of interest, and this discretization is traditionally made of tetrahedral or hexahedral elements. The topology defined by these elements is usually considered as being invariant. However, this is not a realistic assumption if the model is to be employed during a real surgical intervention.

The first objective of this work is to jointly develop advanced topological operations and new finite element approaches that can leverage the use of dynamic topologies. In particular we focus our research on multi-resolution meshes where elements are subdivided in areas where numerical errors need to be kept small [51], [55].

Our second objective is to improve, at the numerical level, the efficiency, robustness, and quality of the simulations. To reach these goals, we essentially rely on two main directions: adaptive meshing to allow mesh transformations during a simulation and support cuts, local remeshing or dynamic refinement in areas of interest; and numerical techniques, such as asynchronous solvers, domain decomposition and model order reduction [35], [36], [46], [47].

We also work on mixed Finite Element Modeling where both tetrahedra and hexahedra can be used at the same time, allowing an ideal compromise between numerical efficiency and mesh adaptation to complex geometries. This research also includes the study of domain decomposition techniques and other coupling techniques for multi-domain multi-physics simulations.

Once the problem, as defined in the previous challenge, has been discretized, we need to solve a large system of linear or nonlinear equations. In both cases, it is necessary to employ numerical solvers repeatedly to construct the solution representing the state of the simulated system. In the past years, we have contributed to this topic through our work on asynchronous preconditioning [36]. We would like to pursue this area of research exploiting the relevant advances in hierarchy-based topologies (e.g. the multi-grid methods). We will also consider advanced non-linear solvers which are necessary for correct resolution of hyper-elastic models and composite models.

Finally, to improve computational times from a programming stand-point, we have started a collaboration with the CAMUS team at Inria. This collaboration aims at using smart code analysis and on-the-fly parallelism to automatically speed-up computation times. In a typical scenario, the modeled organ or tissue is surrounded by its environment represented by other organs, connective tissues or fat. Further, during the intervention, the tissues are manipulated with instruments. Therefore, the interaction will also be an important aspect of our research. We have already developed methods for modeling of advanced interactions between organs, tissues and tools [50], [37]. We will continue exploiting novel methods such as partial factorization [56] and integrate our approach with other techniques such as augmented Lagrangian.

3.3. Data-driven Simulation

Image-driven simulation is a recent area of research that has the potential to bridge the gap between medical imaging and clinical routine by adapting pre-operative data to the time of the procedure. Several challenges are related to image-guided therapy but the main issue consists in aligning pre-operative images onto the patient and keep this alignment up-to-date during the procedure. As most procedures deal with soft-tissues, elastic registration techniques are necessary to perform this step.

Recently, registration techniques started to account for soft tissue biomechanics using physically-based methods, yet several limitations still hinder the use of image-guided therapy in clinical routine. First, as registration methods become more complex, their computation times increase, thus lacking responsiveness. Second, techniques used for non-rigid registration or deformable augmented reality only *borrow* ideas from continuum mechanics but lack some key elements (such as identification of the rest shape, or definition of the boundary conditions). Also, these registration or augmented reality problems are highly dependent on the choice of image modality and require investigating some aspects of computer vision or medical image processing.

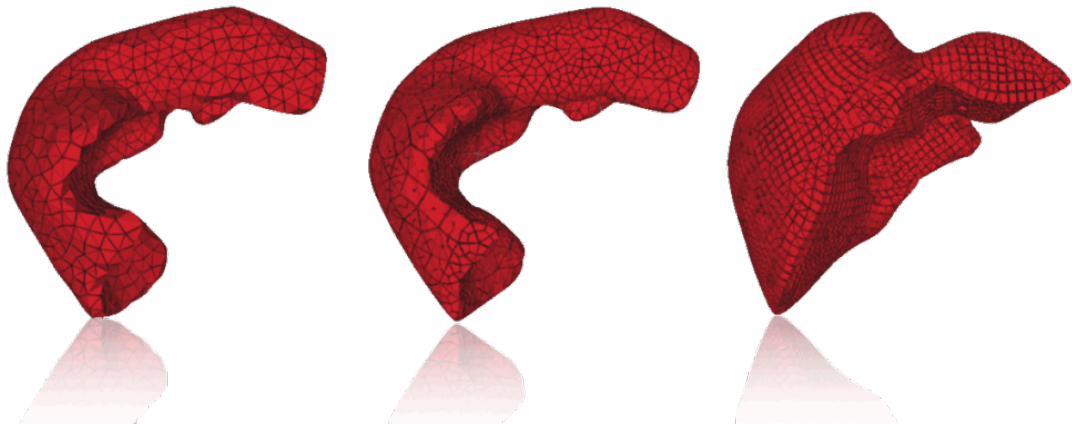


Figure 3. Example of mesh refinement on a complex geometry.

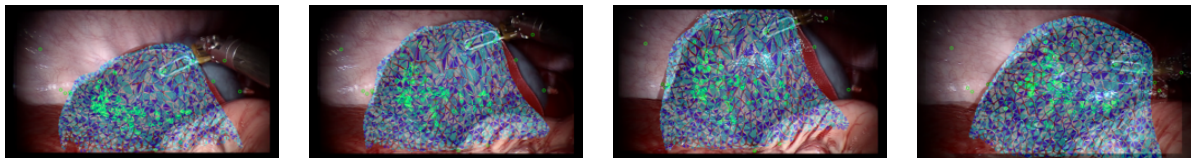


Figure 4. Real-time deformation of a virtual liver according to tissue motion tracked in laparoscopic images.

However, if we can properly address these challenges, the combination of a real-time simulation and regular acquisitions of image data during the procedure opens up very interesting possibilities by using data assimilation to better adapt the model to the intra-operative data [45], [43].

In the area of non-rigid registration and augmented reality, we have already demonstrated the benefit of our physics-based approaches. This was applied in particular to the problem of organ tracking during surgery (Figure 4) and led to several key publications [40], [48], [39] and awards (best paper ISMAR 2013, second best paper at IPCAI 2014). We continue this work with an emphasis on robustness to uncertainty and outliers in the information extracted in real-time from image data and by improving upon our current computer vision techniques, in particular to guarantee a very accurate initial registration of the pre-operative model onto the per-operative surface patch extracted from monocular or stereo laparoscopic cameras. This work will finally benefit from advances in the challenges listed previously, in particular real-time hyper-elastic models of behavior.

The use of simulation in the context of image-guided therapy can be extended in several other ways. A direction we are addressing is the combined use of simulation and X-ray imaging during interventional radiology procedures. Whether it is for percutaneous procedures or catheterization, the task of the simulation is to provide a short-term (1 to 5 seconds) prediction of the needle or catheter position. Using information extracted from the image, the parameters of the simulation can be assimilated (using methods such as unscented Kalman filters [41] and its reduced-order versions [44]), so that the simulation progressively matches the real data in order to reduce uncertainties. We have already started to create a flexible framework integrating the real-time soft-tissue simulation and state-of-the-art methods of data assimilation and filtering [49].

MNEMOSYNE Project-Team

3. Research Program

3.1. Integrative and Cognitive Neuroscience

The human brain is often considered as the most complex system dedicated to information processing. This multi-scale complexity, described from the metabolic to the network level, is particularly studied in integrative neuroscience, the goal of which is to explain how cognitive functions (ranging from sensorimotor coordination to executive functions) emerge from (are the result of the interaction of) distributed and adaptive computations of processing units, displayed along neural structures and information flows. Indeed, beyond the astounding complexity reported in physiological studies, integrative neuroscience aims at extracting, in simplifying models, regularities at various levels of description. From a mesoscopic point of view, most neuronal structures (and particularly some of primary importance like the cortex, cerebellum, striatum, hippocampus) can be described through a regular organization of information flows and homogenous learning rules, whatever the nature of the processed information. From a macroscopic point of view, the arrangement in space of neuronal structures within the cerebral architecture also obeys a functional logic, the sketch of which is captured in models describing the main information flows in the brain, the corresponding loops built in interaction with the external and internal (bodily and hormonal) world and the developmental steps leading to the acquisition of elementary sensorimotor skills up to the most complex executive functions.

In summary, integrative neuroscience builds, on an overwhelming quantity of data, a simplifying and interpretative grid suggesting homogenous local computations and a structured and logical plan for the development of cognitive functions. They arise from interactions and information exchange between neuronal structures and the external and internal world and also within the network of structures.

This domain is today very active and stimulating because it proposes, of course at the price of simplifications, global views of cerebral functioning and more local hypotheses on the role of subsets of neuronal structures in cognition. In the global approaches, the integration of data from experimental psychology and clinical studies leads to an overview of the brain as a set of interacting memories, each devoted to a specific kind of information processing [46]. It results also in longstanding and very ambitious studies for the design of cognitive architectures aiming at embracing the whole cognition. With the notable exception of works initiated by [43], most of these frameworks (e.g. Soar, ACT-R), though sometimes justified on biological grounds, do not go up to a *connectionist* neuronal implementation. Furthermore, because of the complexity of the resulting frameworks, they are restricted to simple symbolic interfaces with the internal and external world and to (relatively) small-sized internal structures. Our main research objective is undoubtedly to build such a general purpose cognitive architecture (to model the brain *as a whole* in a systemic way), using a connectionist implementation and able to cope with a realistic environment.

3.2. Computational Neuroscience

From a general point of view, computational neuroscience can be defined as the development of methods from computer science and applied mathematics, to explore more technically and theoretically the relations between structures and functions in the brain [48], [36]. During the recent years this domain has gained an increasing interest in neuroscience and has become an essential tool for scientific developments in most fields in neuroscience, from the molecule to the system. In this view, all the objectives of our team can be described as possible progresses in computational neuroscience. Accordingly, it can be underlined that the systemic view that we promote can offer original contributions in the sense that, whereas most classical models in computational neuroscience focus on the better understanding of the structure/function relationship for isolated specific structures, we aim at exploring synergies between structures. Consequently, we target interfaces and interplay between heterogeneous modes of computing, which is rarely addressed in classical computational neuroscience.

We also insist on another aspect of computational neuroscience which is, in our opinion, at the core of the involvement of computer scientists and mathematicians in the domain and on which we think we could particularly contribute. Indeed, we think that our primary abilities in numerical sciences imply that our developments are characterized above all by the effectiveness of the corresponding computations: We provide biologically inspired architectures with effective computational properties, such as robustness to noise, self-organization, on-line learning. We more generally underline the requirement that our models must also mimic biology through its most general law of homeostasis and self-adaptability in an unknown and changing environment. This means that we propose to numerically experiment such models and thus provide effective methods to falsify them.

Here, computational neuroscience means mimicking original computations made by the neuronal substratum and mastering their corresponding properties: computations are distributed and adaptive; they are performed without an homunculus or any central clock. Numerical schemes developed for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [33], [42] and were the basis for several contributions in our group [47], [44], [49]. Ensuring such a rigor in the computations associated to our systemic and large scale approach is of central importance.

Equally important is the choice for the formalism of computation, extensively discussed in the connectionist domain. Spiking neurons are today widely recognized of central interest to study synchronization mechanisms and neuronal coupling at the microscopic level [34]; the associated formalism [39] can be possibly considered for local studies or for relating our results with this important domain in connectionism. Nevertheless, we remain mainly at the mesoscopic level of modeling, the level of the neuronal population, and consequently interested in the formalism developed for dynamic neural fields [31], that demonstrated a richness of behavior [35] adapted to the kind of phenomena we wish to manipulate at this level of description. Our group has a long experience in the study and adaptation of the properties of neural fields [44], [45] and their use for observing the emergence of typical cortical properties [38]. In the envisioned development of more complex architectures and interplay between structures, the exploration of mathematical properties such as stability and boundedness and the observation of emerging phenomena is one important objective. This objective is also associated with that of capitalizing our experience and promoting good practices in our software production. In summary, we think that this systemic approach also brings to computational neuroscience new case studies where heterogenous and adaptive models with various time scales and parameters have to be considered jointly to obtain a mastered substratum of computation. This is particularly critical for large scale deployments.

3.3. Machine Learning

The adaptive properties of the nervous system are certainly among its most fascinating characteristics, with a high impact on our cognitive functions. Accordingly, machine learning is a domain [41] that aims at giving such characteristics to artificial systems, using a mathematical framework (probabilities, statistics, data analysis, etc.). Some of its most famous algorithms are directly inspired from neuroscience, at different levels. Connectionist learning algorithms implement, in various neuronal architectures, weight update rules, generally derived from the hebbian rule, performing non supervised (e.g. Kohonen self-organizing maps), supervised (e.g. layered perceptrons) or associative (e.g. Hopfield recurrent network) learning. Other algorithms, not necessarily connectionist, perform other kinds of learning, like reinforcement learning. Machine learning is a very mature domain today and all these algorithms have been extensively studied, at both the theoretical and practical levels, with much success. They have also been related to many functions (in the living and artificial domains) like discrimination, categorisation, sensorimotor coordination, planning, etc. and several neuronal structures have been proposed as the substratum for these kinds of learning [37], [30]. Nevertheless, we believe that, as for previous models, machine learning algorithms remain isolated tools, whereas our systemic approach can bring original views on these problems.

At the cognitive level, most of the problems we face do not rely on only one kind of learning and require instead skills that have to be learned in preliminary steps. That is the reason why cognitive architectures are often referred to as systems of memory, communicating and sharing information for problem solving. Instead of the classical view in machine learning of a flat architecture, a more complex network of modules must be

considered here, as it is the case in the domain of deep learning. In addition, our systemic approach brings the question of incrementally building such a system, with a clear inspiration from developmental sciences. In this perspective, modules can generate internal signals corresponding to internal goals, predictions, error signals, able to supervise the learning of other modules (possibly endowed with a different learning rule), supposed to become autonomous after an instructing period. A typical example is that of episodic learning (in the hippocampus), storing declarative memory about a collection of past episodes and supervising the training of a procedural memory in the cortex.

At the behavioral level, as mentioned above, our systemic approach underlines the fundamental links between the adaptive system and the internal and external world. The internal world includes proprioception and interoception, giving information about the body and its needs for integrity and other fundamental programs. The external world includes physical laws that have to be learned and possibly intelligent agents for more complex interactions. Both involve sensors and actuators that are the interfaces with these worlds and close the loops. Within this rich picture, machine learning generally selects one situation that defines useful sensors and actuators and a corpus with properly segmented data and time, and builds a specific architecture and its corresponding criteria to be satisfied. In our approach however, the first question to be raised is to discover what is the goal, where attention must be focused on and which previous skills must be exploited, with the help of a dynamic architecture and possibly other partners. In this domain, the behavioral and the developmental sciences, observing how and along which stages an agent learns, are of great help to bring some structure to this high dimensional problem.

At the implementation level, this analysis opens many fundamental challenges, hardly considered in machine learning : stability must be preserved despite on-line continuous learning; criteria to be satisfied often refer to behavioral and global measurements but they must be translated to control the local circuit level; in an incremental or developmental approach, how will the development of new functions preserve the integrity and stability of others? In addition, this continuous re-arrangement is supposed to involve several kinds of learning, at different time scales (from msec to years in humans) and to interfere with other phenomena like variability and meta-plasticity.

In summary, our main objective in machine learning is to propose on-line learning systems, where several modes of learning have to collaborate and where the protocols of training are realistic. We promote here a *really autonomous* learning, where the agent must select by itself internal resources (and build them if not available) to evolve at the best in an unknown world, without the help of any *deus-ex-machina* to define parameters, build corpus and define training sessions, as it is generally the case in machine learning. To that end, autonomous robotics (*cf.* § 3.4) is a perfect testbed.

3.4. Autonomous Robotics

Autonomous robots are not only convenient platforms to implement our algorithms; the choice of such platforms is also motivated by theories in cognitive science and neuroscience indicating that cognition emerges from interactions of the body in direct loops with the world (*embodiment of cognition* [32]). In addition to real robotic platforms, software implementations of autonomous robotic systems including components dedicated to their body and their environment will be also possibly exploited, considering that they are also a tool for studying conditions for a real autonomous learning.

A real autonomy can be obtained only if the robot is able to define its goal by itself, without the specification of any high level and abstract cost function or rewarding state. To ensure such a capability, we propose to endow the robot with an artificial physiology, corresponding to perceive some kind of pain and pleasure. It may consequently discriminate internal and external goals (or situations to be avoided). This will mimic circuits related to fundamental needs (e.g. hunger and thirst) and to the preservation of bodily integrity. An important objective is to show that more abstract planning capabilities can arise from these basic goals.

A real autonomy with an on-line continuous learning as described in § 3.3 will be made possible by the elaboration of protocols of learning, as it is the case, in animal conditioning, for experimental studies where performance on a task can be obtained only after a shaping in increasingly complex tasks. Similarly,

developmental sciences can teach us about the ordered elaboration of skills and their association in more complex schemes. An important challenge here is to translate these hints at the level of the cerebral architecture.

As a whole, autonomous robotics permits to assess the consistency of our models in realistic condition of use and offers to our colleagues in behavioral sciences an object of study and comparison, regarding behavioral dynamics emerging from interactions with the environment, also observable at the neuronal level.

In summary, our main contribution in autonomous robotics is to make autonomy possible, by various means corresponding to endow robots with an artificial physiology, to give instructions in a natural and incremental way and to prioritize the synergy between reactive and robust schemes over complex planning structures.

NEUROSYS Project-Team

3. Research Program

3.1. Main Objectives

The main challenge in computational neuroscience is the high complexity of neural systems. The brain is a complex system and exhibits a hierarchy of interacting subunits. On a specific hierarchical level, such subunits evolve on a certain temporal and spatial scale. The interactions of small units on a low hierarchical level build up larger units on a higher hierarchical level evolving on a slower time scale and larger spatial scale. By virtue of the different dynamics on each hierarchical level, until today the corresponding mathematical models and data analysis techniques on each level are still distinct. Only few analysis and modeling frameworks are known which link successfully at least two hierarchical levels.

After extracting models for different description levels, they are typically applied to obtain simulated activity which is supposed to reconstruct features in experimental data. Although this approach appears straightforward, it presents various difficulties. Usually the models involve a large set of unknown parameters which determine the dynamical properties of the models. To optimally reconstruct experimental features, it is necessary to formulate an inverse problem to extract optimally such model parameters from the experimental data. Typically this is a rather difficult problem due to the low signal-to-noise ratio in experimental brain signals. Moreover, the identification of signal features to be reconstructed by the model is not obvious in most applications. Consequently an extended analysis of the experimental data is necessary to identify the interesting data features. It is important to combine such a data analysis step with the parameter extraction procedure to achieve optimal results. Such a procedure depends on the properties of the experimental data and hence has to be developed for each application separately. Machine learning approaches that attempt to mimic the brain and its cognitive processes had a lot of success in classification problems during the last decade. These hierarchical and iterative approaches use non-linear functions, which imitate neural cell responses, to communicate messages between neighboring layers. In our team, we work towards developing polysomnography-specific classifiers that might help in linking the features of particular interest for building systems for sleep signal classification with sleep mechanisms, with the accent on memory consolidation during the Rapid Eye Movement (REM) sleep phase.

3.2. Challenges

Eventually the implementation of the models and analysis techniques achieved promises to be able to construct novel data monitors. This construction involves additional challenges and requires contact with realistic environments. By virtue of the specific applications of the research, the close contact to hospitals and medical enterprises shall be established in a longer term in order to (i) gain deeper insight into the specific application of the devices and (ii) build specific devices in accordance to the actual need. Collaborations with local and national hospitals and the pharmaceutical industry already exist.

3.3. Research Directions

- From the microscopic to the mesoscopic scale:
One research direction focuses on the *relation of single neuron activity on the microscopic scale to the activity of neuronal populations*. To this end, the team investigates the stochastic dynamics of single neurons subject to external random inputs and involving random microscopic properties, such as random synaptic strengths and probability distributions of spatial locations of membrane ion channels. Such an approach yields a stochastic model of single neurons and allows the derivation of a stochastic neural population model.

This bridge between the microscopic and mesoscopic scale may be performed via two pathways. The analytical and numerical treatment of the microscopic model may be called a *bottom-up approach*,

since it leads to a population activity model based on microscopic activity. This approach allows theoretical neural population activity to be compared to experimentally obtained population activity. The *top-down approach* aims at extracting signal features from experimental data gained from neural populations which give insight into the dynamics of neural populations and the underlying microscopic activity. The work on both approaches represents a well-balanced investigation of the neural system based on the systems properties.

- From the mesoscopic to the macroscopic scale:
The other research direction aims to link neural population dynamics to macroscopic activity and behavior or, more generally, to phenomenological features. This link is more indirect but a very powerful approach to understand the brain, e.g., in the context of medical applications. Since real neural systems, such as in mammals, exhibit an interconnected network of neural populations, the team studies analytically and numerically the network dynamics of neural populations to gain deeper insight into possible phenomena, such as traveling waves or enhancement and diminution of certain neural rhythms. Electroencephalography (EEG) is a wonderful brain imaging technique to study the overall brain activity in real time non-invasively. However it is necessary to develop robust techniques based on stable features by investigating the time and frequency domains of brain signals. Two types of information are typically used in EEG signals: (i) transient events such as evoked potentials, spindles and K-complexes and (ii) the power in specific frequency bands.

PARIETAL Project-Team

3. Research Program

3.1. Inverse problems in Neuroimaging

Many problems in neuroimaging can be framed as forward and inverse problems. For instance, brain population imaging is concerned with the *inverse problem* that consists in predicting individual information (behavior, phenotype) from neuroimaging data, while the corresponding *forward problem* boils down to explaining neuroimaging data with the behavioral variables. Solving these problems entails the definition of two terms: a loss that quantifies the goodness of fit of the solution (does the model explain the data well enough?), and a regularization scheme that represents a prior on the expected solution of the problem. These priors can be used to enforce some properties on the solutions, such as sparsity, smoothness or being piece-wise constant.

Let us detail the model used in typical inverse problem: Let \mathbf{X} be a neuroimaging dataset as an $(n_{\text{subjects}}, n_{\text{voxels}})$ matrix, where n_{subjects} and n_{voxels} are the number of subjects under study, and the image size respectively, \mathbf{Y} a set of values that represent characteristics of interest in the observed population, written as $(n_{\text{subjects}}, n_{\text{features}})$ matrix, where n_{features} is the number of characteristics that are tested, and \mathbf{w} an array of shape $(n_{\text{voxels}}, n_{\text{features}})$ that represents a set of pattern-specific maps. In the first place, we may consider the columns $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_{\text{features}}}$ of \mathbf{Y} independently, yielding n_{features} problems to be solved in parallel:

$$\mathbf{Y}_i = \mathbf{X}\mathbf{w}_i + \epsilon_i, \forall i \in \{1, \dots, n_{\text{features}}\},$$

where the vector contains \mathbf{w}_i is the i^{th} row of \mathbf{w} . As the problem is clearly ill-posed, it is naturally handled in a regularized regression framework:

$$\hat{w}_i = \operatorname{argmin}_{\mathbf{w}_i} \|\mathbf{Y}_i - \mathbf{X}\mathbf{w}_i\|^2 + \Psi(\mathbf{w}_i), \quad (13)$$

where Ψ is an adequate penalization used to regularize the solution:

$$\Psi(\mathbf{w}; \lambda_1, \lambda_2, \eta_1, \eta_2) = \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2 + \eta_1 \|\nabla \mathbf{w}\|_{2,1} + \eta_2 \|\nabla \mathbf{w}\|_{2,2} \quad (14)$$

with $\lambda_1, \lambda_2, \eta_1, \eta_2 \geq 0$ (this formulation particularly highlights the fact that convex regularizers are norms or quasi-norms). In general, only one or two of these constraints is considered (hence is enforced with a non-zero coefficient):

- When $\lambda_1 > 0$ only (LASSO), and to some extent, when $\lambda_1, \lambda_2 > 0$ only (elastic net), the optimal solution \mathbf{w} is (possibly very) sparse, but may not exhibit a proper image structure; it does not fit well with the intuitive concept of a brain map.
- Total Variation regularization (see Fig. 1) is obtained for $(\eta_1 > 0)$ only, and typically yields a piece-wise constant solution. It can be associated with Lasso to enforce both sparsity and sparse variations.
- Smooth lasso is obtained with $(\eta_2 > 0 \text{ and } \lambda_1 > 0)$ only, and yields smooth, compactly supported spatial basis functions.

Note that, while the qualitative aspect of the solutions are very different, the predictive power of these models is often very close.

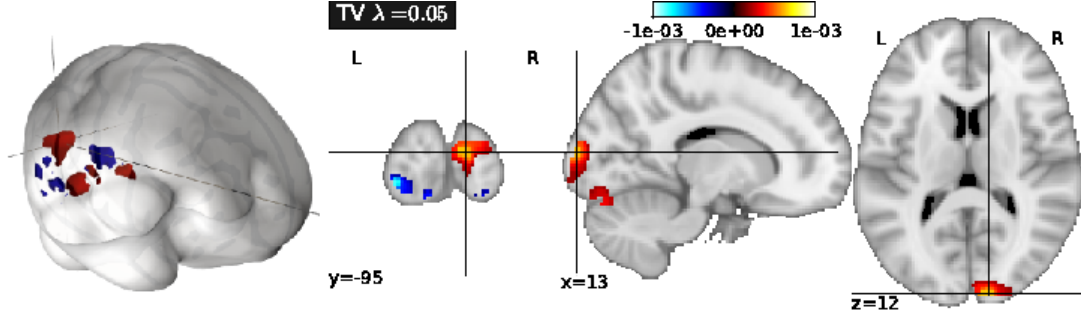


Figure 1. Example of the regularization of a brain map with total variation in an inverse problem. The problem here is to predict the spatial scale of an object presented as a stimulus, given functional neuroimaging data acquired during the presentation of an image. Learning and test are performed across individuals. Unlike other approaches, Total Variation regularization yields a sparse and well-localized solution that also enjoys high predictive accuracy.

The performance of the predictive model can simply be evaluated as the amount of variance in \mathbf{Y}_i fitted by the model, for each $i \in \{1, \dots, n_{features}\}$. This can be computed through cross-validation, by *learning* $\hat{\mathbf{w}}_i$ on some part of the dataset, and then estimating $\|\mathbf{Y}_i - \mathbf{X}\hat{\mathbf{w}}_i\|^2$ using the remainder of the dataset.

This framework is easily extended by considering

- *Grouped penalization*, where the penalization explicitly includes a prior clustering of the features, i.e. voxel-related signals, into given groups. This amounts to enforcing structured priors on the solution.
- *Combined penalizations*, i.e. a mixture of simple and group-wise penalizations, that allow some variability to fit the data in different populations of subjects, while keeping some common constraints.
- *Logistic and hinge regression*, where a non-linearity is applied to the linear model so that it yields a probability of classification in a binary classification problem.
- *Robustness to between-subject variability* to avoid the learned model overly reflecting a few outlying particular observations of the training set. Note that noise and deviating assumptions can be present in both \mathbf{Y} and \mathbf{X} .
- *Multi-task learning*: if several target variables are thought to be related, it might be useful to constrain the estimated parameter vector \mathbf{w} to have a shared support across all these variables. For instance, when one of the variables \mathbf{Y}_i is not well fitted by the model, the estimation of other variables $\mathbf{Y}_j, j \neq i$ may provide constraints on the support of \mathbf{w}_i and thus, improve the prediction of \mathbf{Y}_i .

$$\mathbf{Y} = \mathbf{X}\mathbf{w} + \epsilon, \quad (15)$$

then

$$\hat{\mathbf{w}} = \underset{\mathbf{w}=(\mathbf{w}_i), i=1..n_f}{\operatorname{argmin}} \sum_{i=1}^{n_f} \|\mathbf{Y}_i - \mathbf{X}\mathbf{w}_i\|^2 + \lambda \sum_{j=1}^{n_{voxels}} \sqrt{\sum_{i=1}^{n_f} \mathbf{w}_{i,j}^2} \quad (16)$$

3.2. Multivariate decompositions

Multivariate decompositions provide a way to model complex data such as brain activation images: for instance, one might be interested in extracting an *atlas of brain regions* from a given dataset, such as regions exhibiting similar activity during a protocol, across multiple protocols, or even in the absence of protocol (during resting-state). These data can often be factorized into spatial-temporal components, and thus can be estimated through *regularized Principal Components Analysis* (PCA) algorithms, which share some common steps with regularized regression.

Let \mathbf{X} be a neuroimaging dataset written as an $(n_{subjects}, n_{voxels})$ matrix, after proper centering; the model reads

$$\mathbf{X} = \mathbf{A}\mathbf{D} + \epsilon, \quad (17)$$

where \mathbf{D} represents a set of n_{comp} spatial maps, hence a matrix of shape (n_{comp}, n_{voxels}) , and \mathbf{A} the associated subject-wise loadings. While traditional PCA and independent components analysis are limited to reconstructing components \mathbf{D} within the space spanned by the column of \mathbf{X} , it seems desirable to add some constraints on the rows of \mathbf{D} , that represent spatial maps, such as sparsity, and/or smoothness, as it makes the interpretation of these maps clearer in the context of neuroimaging. This yields the following estimation problem:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{A}\mathbf{D}\|^2 + \Psi(\mathbf{D}) \text{ s.t. } \|\mathbf{A}_i\| = 1 \forall i \in \{1..n_{features}\}, \quad (18)$$

where (\mathbf{A}_i) , $i \in \{1..n_{features}\}$ represents the columns of \mathbf{A} . Ψ can be chosen such as in Eq. (2) in order to enforce smoothness and/or sparsity constraints.

The problem is not jointly convex in all the variables but each penalization given in Eq (2) yields a convex problem on \mathbf{D} for \mathbf{A} fixed, and conversely. This readily suggests an alternate optimization scheme, where \mathbf{D} and \mathbf{A} are estimated in turn, until convergence to a local optimum of the criterion. As in PCA, the extracted components can be ranked according to the amount of fitted variance. Importantly, also, estimated PCA models can be interpreted as a probabilistic model of the data, assuming a high-dimensional Gaussian distribution (probabilistic PCA).

Ultimately, the main limitations to these algorithms is the cost due to the memory requirements: holding datasets with large dimension and large number of samples (as in recent neuroimaging cohorts) leads to inefficient computation. To solve this issue, online methods are particularly attractive [29].

3.3. Covariance estimation

Another important estimation problem stems from the general issue of learning the relationship between sets of variables, in particular their covariance. Covariance learning is essential to model the dependence of these variables when they are used in a multivariate model, for instance to study potential interactions among them and with other variables. Covariance learning is necessary to model latent interactions in high-dimensional observation spaces, e.g. when considering multiple contrasts or functional connectivity data.

The difficulties are two-fold: on the one hand, there is a shortage of data to learn a good covariance model from an individual subject, and on the other hand, subject-to-subject variability poses a serious challenge to the use of multi-subject data. While the covariance structure may vary from population to population, or depending on the input data (activation versus spontaneous activity), assuming some shared structure across problems, such as their sparsity pattern, is important in order to obtain correct estimates from noisy data. Some of the most important models are:

- **Sparse Gaussian graphical models**, as they express meaningful conditional independence relationships between regions, and do improve conditioning/avoid overfit.

- **Decomposable models**, as they enjoy good computational properties and enable intuitive interpretations of the network structure. Whether they can faithfully or not represent brain networks is still an open question.
- **PCA-based regularization of covariance** which is powerful when modes of variation are more important than conditional independence relationships.

Adequate model selection procedures are necessary to achieve the right level of sparsity or regularization in covariance estimation; the natural evaluation metric here is the out-of-sample likelihood of the associated Gaussian model. Another essential remaining issue is to develop an adequate statistical framework to test differences between covariance models in different populations. To do so, we consider different means of parametrizing covariance distributions and how these parametrizations impact the test of statistical differences across individuals.

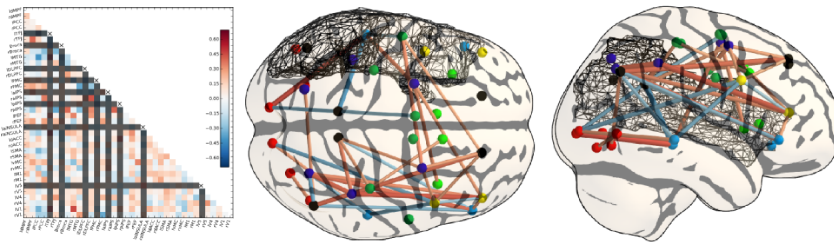


Figure 2. Example of functional connectivity analysis: The correlation matrix describing brain functional connectivity in a post-stroke patient (lesion volume outlined as a mesh) is compared to a group of control subjects. Some edges of the graphical model show a significant difference, but the statistical detection of the difference requires a sophisticated statistical framework for the comparison of graphical models.

VISAGES Project-Team

3. Research Program

3.1. Research Program

The scientific foundations of our team concern the development of new processing algorithms in the field of medical image computing : image fusion (registration and visualization), image segmentation and analysis, management of image related information. Since this is a very large domain, which can be applied on numerous types of application; for seek of efficiency, the purpose of our methodological work primarily focuses on clinical aspects and for the most part on head and neck related diseases. In addition, we emphasize our research efforts on the neuroimaging domain. Concerning the scientific foundations, we have pushed our research efforts:

- In the field of image fusion and image registration (rigid and deformable transformations) with a special emphasis on new challenging registration issues, especially when statistical approaches based on joint histogram cannot be used or when the registration stage has to cope with loss or appearance of material (like in surgery or in tumor imaging for instance).
- In the field of image analysis and statistical modeling with a new focus on image feature and group analysis problems. A special attention was also to develop advanced frameworks for the construction of atlases and for automatic and supervised labeling of brain structures.
- In the field of image segmentation and structure recognition, with a special emphasis on the difficult problems of *i*) image restoration for new imaging sequences (new Magnetic Resonance Imaging protocols, 3D ultrasound sequences...), and *ii*) structure segmentation and labelling based on shape, multimodal and statistical information.
- Following past national projects where we had leading roles (e.g., Neurobase, NeuroLog, . . .), we wanted to enhance the development of distributed and heterogeneous medical image processing systems.

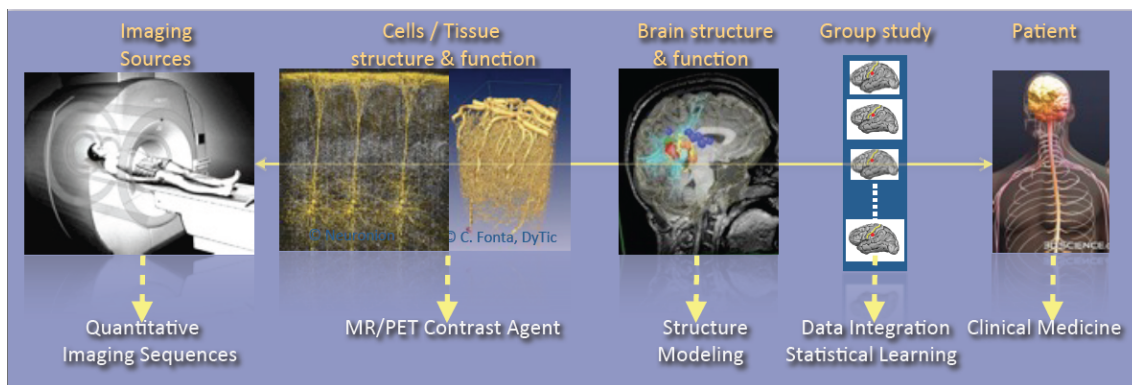


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales: from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions, to the population level for the modelling of group patterns and the learning of group or individual imaging markers.

As shown in Fig. 1, research activities of the VISAGES U1228 team are tightly coupling observations and models through integration of clinical and multi-scale data, phenotypes (cellular, molecular or structural patterns). We work on personalized models of central nervous system organs and pathologies, and intend to confront these models to clinical investigation studies for quantitative diagnosis, prevention of diseases, therapy planning and validation. These approaches are developed in a translational framework where the data integration process to build the models inherits from specific clinical studies, and where the models are assessed on prospective clinical trials for diagnosis and therapy planning. All of this research activity is conducted in tight links with the **Neurinfo** imaging platform environments and the engineering staff of the platform. In this context, some of our major challenges in this domain concern:

- The elaboration of new descriptors to study the brain structure and function (e.g., variation of brain perfusion with and without contrast agent, evolution in shape and size of an anatomical structure in relation with normal, pathological or functional patterns, computation of asymmetries from shapes and volumes).
- The integration of additional spatio-temporal imaging sequences covering a larger range of observation, from the molecular level to the organ through the cell (Arterial Spin Labeling, diffusion MRI, MR relaxometry, MR cell labeling imaging, PET molecular imaging, ...). This includes the elaboration of new image descriptors coming from spatio-temporal quantitative or contrast-enhanced MRI.
- The creation of computational models through data fusion of molecular, cellular, structural and functional image descriptors from group studies of normal and/or pathological subjects.
- The evaluation of these models on acute pathologies especially for the study of degenerative, psychiatric or developmental brain diseases (e.g., Multiple Sclerosis, Epilepsy, Parkinson, Dementia, Strokes, Depression, Schizophrenia, ...) in a translational framework.

In terms of methodological developments, we are particularly working on statistical methods for multidimensional image analysis, and feature selection and discovery, which include:

- The development of specific shape and appearance models, construction of atlases better adapted to a patient or a group of patients in order to better characterize the pathology;
- The development of advanced segmentation and modeling methods dealing with longitudinal and multidimensional data (vector or tensor fields), especially with the integration of new prior models to control the integration of multiscale data and aggregation of models;
- The development of new models and probabilistic methods to create water diffusion maps from MRI;
- The integration of machine learning procedures for classification and labeling of multidimensional features (from scalar to tensor fields and/or geometric features): pattern and rule inference and knowledge extraction are key techniques to help in the elaboration of knowledge in the complex domains we address;
- The development of new dimensionality reduction techniques for problems with massive data, which includes dictionary learning for sparse model discovery. Efficient techniques have still to be developed to properly extract from a raw mass of images derived data that are easier to analyze.

AIRSEA Project-Team

3. Research Program

3.1. Introduction

Recent events have raised questions regarding the social and economic implications of anthropic alterations of the Earth system, i.e. climate change and the associated risks of increasing extreme events. Ocean and atmosphere, coupled with other components (continent and ice) are the building blocks of the Earth system. A better understanding of the ocean atmosphere system is a key ingredient for improving prediction of such events. Numerical models are essential tools to understand processes, and simulate and forecast events at various space and time scales. Geophysical flows generally have a number of characteristics that make it difficult to model them. This justifies the development of specifically adapted mathematical methods:

- Geophysical flows are strongly non-linear. Therefore, they exhibit interactions between different scales, and unresolved small scales (smaller than mesh size) of the flows have to be **parameterized** in the equations.
- Geophysical fluids are non closed systems. They are open-ended in their scope for including and dynamically coupling different physical processes (e.g., atmosphere, ocean, continental water, etc). **Coupling** algorithms are thus of primary importance to account for potentially significant feedback.
- Numerical models contain parameters which cannot be estimated accurately either because they are difficult to measure or because they represent some poorly known subgrid phenomena. There is thus a need for **dealing with uncertainties**. This is further complicated by the turbulent nature of geophysical fluids.
- The computational cost of geophysical flow simulations is huge, thus requiring the use of **reduced models, multiscale methods** and the design of algorithms ready for **high performance computing** platforms.

Our scientific objectives are divided into four major points. The first objective focuses on developing advanced mathematical methods for both the ocean and atmosphere, and the coupling of these two components. The second objective is to investigate the derivation and use of model reduction to face problems associated with the numerical cost of our applications. The third objective is directed toward the management of uncertainty in numerical simulations. The last objective deals with efficient numerical algorithms for new computing platforms. As mentioned above, the targeted applications cover oceanic and atmospheric modeling and related extreme events using a hierarchy of models of increasing complexity.

3.2. Modeling for oceanic and atmospheric flows

Current numerical oceanic and atmospheric models suffer from a number of well-identified problems. These problems are mainly related to lack of horizontal and vertical resolution, thus requiring the parameterization of unresolved (subgrid scale) processes and control of discretization errors in order to fulfill criteria related to the particular underlying physics of rotating and strongly stratified flows. Oceanic and atmospheric coupled models are increasingly used in a wide range of applications from global to regional scales. Assessment of the reliability of those coupled models is an emerging topic as the spread among the solutions of existing models (e.g., for climate change predictions) has not been reduced with the new generation models when compared to the older ones.

Advanced methods for modeling 3D rotating and stratified flows The continuous increase of computational power and the resulting finer grid resolutions have triggered a recent regain of interest in numerical methods and their relation to physical processes. Going beyond present knowledge requires a better understanding of numerical dispersion/dissipation ranges and their connection to model fine scales. Removing the leading order truncation error of numerical schemes is thus an active topic of research and each mathematical tool has to adapt to the characteristics of three dimensional stratified and rotating flows. Studying the link between discretization errors and subgrid scale parameterizations is also arguably one of the main challenges.

Complexity of the geometry, boundary layers, strong stratification and lack of resolution are the main sources of discretization errors in the numerical simulation of geophysical flows. This emphasizes the importance of the definition of the computational grids (and coordinate systems) both in horizontal and vertical directions, and the necessity of truly multi resolution approaches. At the same time, the role of the small scale dynamics on large scale circulation has to be taken into account. Such parameterizations may be of deterministic as well as stochastic nature and both approaches are taken by the AIRSEA team. The design of numerical schemes consistent with the parameterizations is also arguably one of the main challenges for the coming years. This work is complementary and linked to that on parameters estimation described in 3.4 .

Ocean Atmosphere interactions and formulation of coupled models State-of-the-art climate models (CMs) are complex systems under continuous development. A fundamental aspect of climate modeling is the representation of air-sea interactions. This covers a large range of issues: parameterizations of atmospheric and oceanic boundary layers, estimation of air-sea fluxes, time-space numerical schemes, non conforming grids, coupling algorithms ...Many developments related to these different aspects were performed over the last 10-15 years, but were in general conducted independently of each other.

The aim of our work is to revisit and enrich several aspects of the representation of air-sea interactions in CMs, paying special attention to their overall consistency with appropriate mathematical tools. We intend to work consistently on the physics and numerics. Using the theoretical framework of global-in-time Schwarz methods, our aim is to analyze the mathematical formulation of the parameterizations in a coupling perspective. From this study, we expect improved predictability in coupled models (this aspect will be studied using techniques described in 3.4). Complementary work on space-time nonconformities and acceleration of convergence of Schwarz-like iterative methods (see 7.1.2) are also conducted.

3.3. Model reduction / multiscale algorithms

The high computational cost of the applications is a common and major concern to have in mind when deriving new methodological approaches. This cost increases dramatically with the use of sensitivity analysis or parameter estimation methods, and more generally with methods that require a potentially large number of model integrations.

A dimension reduction, using either stochastic or deterministic methods, is a way to reduce significantly the number of degrees of freedom, and therefore the calculation time, of a numerical model.

Model reduction Reduction methods can be deterministic (proper orthogonal decomposition, other reduced bases) or stochastic (polynomial chaos, Gaussian processes, kriging), and both fields of research are very active. Choosing one method over another strongly depends on the targeted application, which can be as varied as real-time computation, sensitivity analysis (see e.g., section 7.3.1) or optimisation for parameter estimation (see below).

Our goals are multiple, but they share a common need for certified error bounds on the output. Our team has a 4-year history of working on certified reduction methods and has a unique positioning at the interface between deterministic and stochastic approaches. Thus, it seems interesting to conduct a thorough comparison of the two alternatives in the context of sensitivity analysis. Efforts will also be directed toward the development of efficient greedy algorithms for the reduction, and the derivation of goal-oriented sharp error bounds for non linear models and/or non linear outputs of interest. This will be complementary to our work on the deterministic reduction of parametrized viscous Burgers and Shallow Water equations where the objective is to obtain sharp error bounds to provide confidence intervals for the estimation of sensitivity indices.

Reduced models for coupling applications Global and regional high-resolution oceanic models are either coupled to an atmospheric model or forced at the air-sea interface by fluxes computed empirically preventing proper physical feedback between the two media. Thanks to high-resolution observational studies, the existence of air-sea interactions at oceanic mesoscales (i.e., at $\mathcal{O}(1km)$ scales) have been unambiguously shown. Those interactions can be represented in coupled models only if the oceanic and atmospheric models are run on the same high-resolution computational grid, and are absent in a forced mode. Fully coupled models

at high-resolution are seldom used because of their prohibitive computational cost. The derivation of a reduced model as an alternative between a forced mode and the use of a full atmospheric model is an open problem.

Multiphysics coupling often requires iterative methods to obtain a mathematically correct numerical solution. To mitigate the cost of the iterations, we will investigate the possibility of using reduced-order models for the iterative process. We will consider different ways of deriving a reduced model: coarsening of the resolution, degradation of the physics and/or numerical schemes, or simplification of the governing equations. At a mathematical level, we will strive to study the well-posedness and the convergence properties when reduced models are used. Indeed, running an atmospheric model at the same resolution as the ocean model is generally too expensive to be manageable, even for moderate resolution applications. To account for important fine-scale interactions in the computation of the air-sea boundary condition, the objective is to derive a simplified boundary layer model that is able to represent important 3D turbulent features in the marine atmospheric boundary layer.

Reduced models for multiscale optimization The field of multigrid methods for optimisation has known a tremendous development over the past few decades. However, it has not been applied to oceanic and atmospheric problems apart from some crude (non-converging) approximations or applications to simplified and low dimensional models. This is mainly due to the high complexity of such models and to the difficulty in handling several grids at the same time. Moreover, due to complex boundaries and physical phenomena, the grid interactions and transfer operators are not trivial to define.

Multigrid solvers (or multigrid preconditioners) are efficient methods for the solution of variational data assimilation problems. We would like to take advantage of these methods to tackle the optimization problem in high dimensional space. High dimensional control space is obtained when dealing with parameter fields estimation, or with control of the full 4D (space time) trajectory. It is important since it enables us to take into account model errors. In that case, multigrid methods can be used to solve the large scales of the problem at a lower cost, this being potentially coupled with a scale decomposition of the variables themselves.

3.4. Dealing with uncertainties

There are many sources of uncertainties in numerical models. They are due to imperfect external forcing, poorly known parameters, missing physics and discretization errors. Studying these uncertainties and their impact on the simulations is a challenge, mostly because of the high dimensionality and non-linear nature of the systems. To deal with these uncertainties we work on three axes of research, which are linked: sensitivity analysis, parameter estimation and risk assessment. They are based on either stochastic or deterministic methods.

Sensitivity analysis Sensitivity analysis (SA), which links uncertainty in the model inputs to uncertainty in the model outputs, is a powerful tool for model design and validation. First, it can be a pre-stage for parameter estimation (see 3.4), allowing for the selection of the more significant parameters. Second, SA permits understanding and quantifying (possibly non-linear) interactions induced by the different processes defining e.g., realistic ocean atmosphere models. Finally SA allows for validation of models, checking that the estimated sensitivities are consistent with what is expected by the theory. On ocean, atmosphere and coupled systems, only first order deterministic SA are performed, neglecting the initialization process (data assimilation). AIRSEA members and collaborators proposed to use second order information to provide consistent sensitivity measures, but so far it has only been applied to simple academic systems. Metamodels are now commonly used, due to the cost induced by each evaluation of complex numerical models: mostly Gaussian processes, whose probabilistic framework allows for the development of specific adaptive designs, and polynomial chaos not only in the context of intrusive Galerkin approaches but also in a black-box approach. Until recently, global SA was based primarily on a set of engineering practices. New mathematical and methodological developments have led to the numerical computation of Sobol' indices, with confidence intervals assessing for both metamodel and estimation errors. Approaches have also been extended to the case of dependent entries, functional inputs and/or output and stochastic numerical codes. Other types of indices and generalizations of Sobol' indices have also been introduced.

Concerning the stochastic approach to SA we plan to work with parameters that show spatio-temporal dependencies and to continue toward more realistic applications where the input space is of huge dimension with highly correlated components. Sensitivity analysis for dependent inputs also introduces new challenges. In our applicative context, it would seem prudent to carefully learn the spatio-temporal dependences before running a global SA. In the deterministic framework we focus on second order approaches where the sought sensitivities are related to the optimality system rather than to the model; i.e., we consider the whole forecasting system (model plus initialization through data assimilation).

All these methods allow for computing sensitivities and more importantly a posteriori error statistics.

Parameter estimation Advanced parameter estimation methods are barely used in ocean, atmosphere and coupled systems, mostly due to a difficulty of deriving adequate response functions, a lack of knowledge of these methods in the ocean-atmosphere community, and also to the huge associated computing costs. In the presence of strong uncertainties on the model but also on parameter values, simulation and inference are closely associated. Filtering for data assimilation and Approximate Bayesian Computation (ABC) are two examples of such association.

Stochastic approach can be compared with the deterministic approach, which allows to determine the sensitivity of the flow to parameters and optimize their values relying on data assimilation. This approach is already shown to be capable of selecting a reduced space of the most influent parameters in the local parameter space and to adapt their values in view of correcting errors committed by the numerical approximation. This approach assumes the use of automatic differentiation of the source code with respect to the model parameters, and optimization of the obtained raw code.

AIRSEA assembles all the required expertise to tackle these difficulties. As mentioned previously, the choice of parameterization schemes and their tuning has a significant impact on the result of model simulations. Our research will focus on parameter estimation for parameterized Partial Differential Equations (PDEs) and also for parameterized Stochastic Differential Equations (SDEs). Deterministic approaches are based on optimal control methods and are local in the parameter space (i.e., the result depends on the starting point of the estimation) but thanks to adjoint methods they can cope with a large number of unknowns that can also vary in space and time. Multiscale optimization techniques as described in 7.2 will be one of the tools used. This in turn can be used either to propose a better (and smaller) parameter set or as a criterion for discriminating parameterization schemes. Statistical methods are global in the parameter state but may suffer from the curse of dimensionality. However, the notion of parameter can also be extended to functional parameters. We may consider as parameter a functional entity such as a boundary condition on time, or a probability density function in a stationary regime. For these purposes, non-parametric estimation will also be considered as an alternative.

Risk assessment Risk assessment in the multivariate setting suffers from a lack of consensus on the choice of indicators. Moreover, once the indicators are designed, it still remains to develop estimation procedures, efficient even for high risk levels. Recent developments for the assessment of financial risk have to be considered with caution as methods may differ pertaining to general financial decisions or environmental risk assessment. Modeling and quantifying uncertainties related to extreme events is of central interest in environmental sciences. In relation to our scientific targets, risk assessment is very important in several areas: hydrological extreme events, cyclone intensity, storm surges...Environmental risks most of the time involve several aspects which are often correlated. Moreover, even in the ideal case where the focus is on a single risk source, we have to face the temporal and spatial nature of environmental extreme events. The study of extremes within a spatio-temporal framework remains an emerging field where the development of adapted statistical methods could lead to major progress in terms of geophysical understanding and risk assessment thus coupling data and model information for risk assessment.

Based on the above considerations we aim to answer the following scientific questions: how to measure risk in a multivariate/spatial framework? How to estimate risk in a non stationary context? How to reduce dimension (see 3.3) for a better estimation of spatial risk?

Extreme events are rare, which means there is little data available to make inferences of risk measures. Risk assessment based on observation therefore relies on multivariate extreme value theory. Interacting particle systems for the analysis of rare events is commonly used in the community of computer experiments. An open question is the pertinence of such tools for the evaluation of environmental risk.

Most numerical models are unable to accurately reproduce extreme events. There is therefore a real need to develop efficient assimilation methods for the coupling of numerical models and extreme data.

3.5. High performance computing

Methods for sensitivity analysis, parameter estimation and risk assessment are extremely costly due to the necessary number of model evaluations. This number of simulations require considerable computational resources, depends on the complexity of the application, the number of input variables and desired quality of approximations. To this aim, the AIRSEA team is an intensive user of HPC computing platforms, particularly grid computing platforms. The associated grid deployment has to take into account the scheduling of a huge number of computational requests and the links with data-management between these requests, all of these as automatically as possible. In addition, there is an increasing need to propose efficient numerical algorithms specifically designed for new (or future) computing architectures and this is part of our scientific objectives. According to the computational cost of our applications, the evolution of high performance computing platforms has to be taken into account for several reasons. While our applications are able to exploit space parallelism to its full extent (oceanic and atmospheric models are traditionally based on a spatial domain decomposition method), the spatial discretization step size limits the efficiency of traditional parallel methods. Thus the inherent parallelism is modest, particularly for the case of relative coarse resolution but with very long integration time (e.g., climate modeling). Paths toward new programming paradigms are thus needed. As a step in that direction, we plan to focus our research on parallel in time methods.

New numerical algorithms for high performance computing Parallel in time methods can be classified into three main groups. In the first group, we find methods using parallelism across the method, such as parallel integrators for ordinary differential equations. The second group considers parallelism across the problem. Falling into this category are methods such as waveform relaxation where the space-time system is decomposed into a set of subsystems which can then be solved independently using some form of relaxation techniques or multigrid reduction in time. The third group of methods focuses on parallelism across the steps. One of the best known algorithms in this family is parareal. Other methods combining the strengths of those listed above (e.g., PFASST) are currently under investigation in the community.

Parallel in time methods are iterative methods that may require a large number of iteration before convergence. Our first focus will be on the convergence analysis of parallel in time (Parareal / Schwarz) methods for the equation systems of oceanic and atmospheric models. Our second objective will be on the construction of fast (approximate) integrators for these systems. This part is naturally linked to the model reduction methods of section (7.2.1). Fast approximate integrators are required both in the Schwarz algorithm (where a first guess of the boundary conditions is required) and in the Parareal algorithm (where the fast integrator is used to connect the different time windows). Our main application of these methods will be on climate (i.e., very long time) simulations. Our second application of parallel in time methods will be in the context of optimization methods. In fact, one of the major drawbacks of the optimal control techniques used in 3.4 is a lack of intrinsic parallelism in comparison with ensemble methods. Here, parallel in time methods also offer ways to better efficiency. The mathematical key point is centered on how to efficiently couple two iterative methods (i.e., parallel in time and optimization methods).

ANGE Project-Team

3. Research Program

3.1. Overview

The research activities carried out within the ANGE team strongly couple the development of methodological tools with applications to real-life problems and the transfer of numerical codes. The main purpose is to obtain new models adapted to the physical phenomena at stake, identify the main properties that reflect the physical meaning of the models (uniqueness, conservativity, entropy dissipation, ...), propose effective numerical methods to approximate their solution in complex configurations (multi-dimensional, unstructured meshes, well-balanced, ...) and to assess the results with data in the purpose of potentially correcting the models.

The difficulties arising in gravity driven flow studies are threefold.

- Models and equations encountered in fluid mechanics (typically the free surface Navier-Stokes equations) are complex to analyze and solve.
- The underlying phenomena often take place over large domains with very heterogeneous length scales (size of the domain, mean depth, wave length, ...) and distinct time scales, *e.g.* coastal erosion, propagation of a tsunami, ...
- These problems are multi-physics with strong couplings and nonlinearities.

3.2. Modelling and analysis

Hazardous flows are complex physical phenomena that can hardly be represented by shallow water type systems of partial differential equations (PDEs). In this domain, the research program is devoted to the derivation and analysis of reduced complexity models compared to the Navier-Stokes equations, but relaxing the shallow water assumptions. The main purpose is then to obtain models well-adapted to the physical phenomena at stake.

Even if the resulting models do not strictly belong to the family of hyperbolic systems, they exhibit hyperbolic features: the analysis and discretisation techniques we intend to develop have connections with those used for hyperbolic conservation laws. It is worth noticing that the need for robust and efficient numerical procedures is reinforced by the smallness of dissipative effects in geophysical models which therefore generate singular solutions and instabilities.

On the one hand, the derivation of the Saint-Venant system from the Navier-Stokes equations is based on two approximations (the so-called shallow water assumptions), namely

- the horizontal fluid velocity is well approximated by its mean value along the vertical direction,
- the pressure is hydrostatic or equivalently the vertical acceleration of the fluid can be neglected compared to the gravitational effects.

As a consequence the objective is to get rid of these two assumptions, one after the other, in order to obtain models accurately approximating the incompressible Euler or Navier-Stokes equations.

On the other hand, many applications require the coupling with non-hydrodynamic equations, as in the case of micro-algae production or erosion processes. These new equations comprise non-hyperbolic features and a special analysis is needed.

3.2.1. Multilayer approach

As for the first shallow water assumption, *multi-layer* systems were proposed to describe the flow as a superposition of Saint-Venant type systems [34], [36], [37]. Even if this approach has provided interesting results, layers are considered separate and non-miscible fluids, which implies strong limitations. That is why we proposed a slightly different approach [1], [2] based on a Galerkin type decomposition along the vertical axis of all variables and leading, both for the model and its discretisation, to more accurate results.

A kinetic representation of our multilayer model allows to derive robust numerical schemes endowed with crucial properties such as: consistency, conservativity, positivity, preservation of equilibria, ... It is one of the major achievements of the team but it needs to be analyzed and extended in several directions namely:

- The convergence of the multilayer system towards the hydrostatic Euler system as the number of layers goes to infinity is a critical point. It is not fully satisfactory to have only formal estimates of the convergence and sharp estimates would provide an optimal number of layers.
- The introduction of several source terms due for instance to the Coriolis force or extra terms from changes of coordinates seems necessary. Their inclusion should lead to substantial modifications of the numerical scheme.
- Its hyperbolicity has not yet been proven and conversely the possible loss of hyperbolicity cannot be characterised. Similarly, the hyperbolic feature is essential in the propagation and generation of waves.

3.2.2. *Non-hydrostatic models*

The hydrostatic assumption consists in neglecting the vertical acceleration of the fluid. It is considered valid for a large class of geophysical flows but is restrictive in various situations where the dispersive effects (like wave propagation) cannot be neglected. For instance, when a wave reaches the coast, bathymetry variations give a vertical acceleration to the fluid that strongly modifies the wave characteristics and especially its height.

Processing an asymptotic expansion (w.r.t. the aspect ratio for shallow water flows) into the Navier-Stokes equations, we obtain at the leading order the Saint-Venant system. Going one step further leads to a vertically averaged version of the Euler/Navier-Stokes equations involving some non-hydrostatic terms. This model has several advantages:

- it admits an energy balance law (that is not the case for most dispersive models available in the literature),
- it reduces to the Saint-Venant system when the non-hydrostatic pressure term vanishes,
- it consists in a set of conservation laws with source terms,
- it does not contain high order derivatives.

3.2.3. *Multi-physics modelling*

The coupling of hydrodynamic equations with other equations in order to model interactions between complex systems represents an important part of the team research. More precisely, three multi-physics systems are investigated. More details about the industrial impact of these studies are presented in the following section.

- To estimate the risk for infrastructures in coastal zones or close to a river, the resolution of the shallow water equations with moving bathymetry is necessary. The first step consisted in the study of an additional equation largely used in engineering science: The Exner equation. The analysis enabled to exhibit drawbacks of the coupled model such as the lack of energy conservation or the strong variations of the solution from small perturbations. A new formulation is proposed to avoid these drawbacks. The new model consists in a coupling between conservation laws and an elliptic equation, like the Euler/Poisson system, suggesting to use well-known strategies for the analysis and the numerical resolution. In addition, the new formulation is derived from classical complex rheology models and allowed physical phenomena like threshold laws.
- Interaction between flows and floating structures is the challenge at the scale of the shallow water equations. This study requires a better understanding of the energy exchanges between the flow and the structure. The mathematical model of floating structures is very hard to solve numerically due to the non-penetration condition at the interface between the flow and the structure. It leads to infinite potential wave speeds that could not be solved with classical free surface numerical schemes. A relaxation model was derived to overcome this difficulty. It represents the interaction with the floating structure with a free surface model-type.

- If the interactions between hydrodynamics and biology phenomena are known through laboratory experiments, it is more difficult to predict the evolution, especially for the biological quantities, in a real and heterogeneous system. The objective is to model and reproduce the hydrodynamics modifications due to forcing term variations (in time and space). We are typically interested in phenomena such as eutrophication, development of harmful bacteria (cyanobacteria) and upwelling phenomena.

3.2.4. Data assimilation and inverse modelling

In environmental applications, the most accurate numerical models remain subject to uncertainties that originate from their parameters and shortcomings in their physical formulations. It is often desirable to quantify the resulting uncertainties in a model forecast. The propagation of the uncertainties may require the generation of ensembles of simulations that ideally sample from the probability density function of the forecast variables. Classical approaches rely on multiple models and on Monte Carlo simulations. The applied perturbations need to be calibrated for the ensemble of simulations to properly sample the uncertainties. Calibrations involve ensemble scores that compare the consistency between the ensemble simulations and the observational data. The computational requirements are so high that designing fast surrogate models or metamodels is often required.

In order to reduce the uncertainties, the fixed or mobile observations of various origins and accuracies can be merged with the simulation results. The uncertainties in the observations and their representativeness also need to be quantified in the process. The assimilation strategy can be formulated in terms of state estimation or parameter estimation (also called inverse modelling). Different algorithms are employed for static and dynamic models, for analyses and forecasts. A challenging question lies in the optimization of the observational network for the assimilation to be the most efficient at a given observational cost.

3.3. Numerical analysis

3.3.1. Non-hydrostatic scheme

The main challenge in the study of the non-hydrostatic model is to design a robust and efficient numerical scheme endowed with properties such as: positivity, wet/dry interfaces treatment, consistency. It must be noticed that even if the non-hydrostatic model looks like an extension of the Saint-Venant system, most of the known techniques used in the hydrostatic case are not efficient as we recover strong difficulties encountered in incompressible fluid mechanics due to the extra pressure term. These difficulties are reinforced by the absence of viscous/dissipative terms.

3.3.2. Space decomposition and adaptive scheme

In the quest for a better balance between accuracy and efficiency, a strategy consists in the adaptation of models. Indeed, the systems of partial differential equations we consider result from a hierarchy of simplifying assumptions. However, some of these hypotheses may turn out to be irrelevant locally. The adaptation of models thus consists in determining areas where a simplified model (*e.g.* shallow water type) is valid and where it is not. In the latter case, we may go back to the “parent” model (*e.g.* Euler) in the corresponding area. This implies to know how to handle the coupling between the aforementioned models from both theoretical and numerical points of view. In particular, the numerical treatment of transmission conditions is a key point. It requires the estimation of characteristic values (Riemann invariant) which have to be determined according to the regime (torrential or fluvial).

3.3.3. Asymptotic-Preserving scheme for source terms

Hydrodynamic models comprise advection and sources terms. The conservation of the balance between source terms, typically viscosity and friction, has a significant impact since the overall flow is generally a perturbation around an equilibrium. The design of numerical schemes able to preserve such balances is a challenge from both theoretical and industrial points of view. The concept of Asymptotic-Preserving (AP) methods is of great interest in order to overcome these issues.

Another difficulty occurs when a term, typically related to the pressure, becomes very large compared to the order of magnitude of the velocity. At this regime, namely the so-called *low Froude* (shallow water) or *low Mach* (Euler) regimes, the difference between the speed of the gravity waves and the physical velocity makes classical numerical schemes inefficient: firstly because of the error of truncation which is inversely proportional to the small parameters, secondly because of the time step governed by the largest speed of the gravity wave. AP methods made a breakthrough in the numerical resolution of asymptotic perturbations of partial-differential equations concerning the first point. The second one can be fixed using partially implicit scheme.

3.3.4. *Multi-physics models*

Coupling problems also arise within the fluid when it contains pollutants, density variations or biological species. For most situations, the interactions are small enough to use a splitting strategy and the classical numerical scheme for each sub-model, whether it be hydrodynamic or non-hydrodynamic.

The sediment transport raises interesting issues from a numerical aspect. This is an example of coupling between the flow and another phenomenon, namely the deformation of the bottom of the basin that can be carried out either by bed load where the sediment has its own velocity or suspended load in which the particles are mostly driven by the flow. This phenomenon involves different time scales and nonlinear retroactions; hence the need for accurate mechanical models and very robust numerical methods. In collaboration with industrial partners (EDF-LNHE), the team already works on the improvement of numerical methods for existing (mostly empirical) models but our aim is also to propose new (quite) simple models that contain important features and satisfy some basic mechanical requirements. The extension of our 3D models to the transport of weighted particles can also be here of great interest.

3.3.5. *Optimisation*

Numerical simulations are a very useful tool for the design of new processes, for instance in renewable energy or water decontamination. The optimisation of the process according to a well-defined objective such as the production of energy or the evaluation of a pollutant concentration is the logical upcoming challenge in order to propose competitive solutions in industrial context. First of all, the set of parameters that have a significant impact on the result and on which we can act in practice is identified. Then the optimal parameters can be obtained using the numerical codes produced by the team to estimate the performance for a given set of parameters with an additional loop such as gradient descent or Monte Carlo method. The optimisation is used in practice to determine the best profile for turbine pales, the best location for water turbine implantation, in particular for a farm.

CASTOR Project-Team

3. Research Program

3.1. Plasma Physics

Participants: Jacques Blum, Cédric Boulbe, Blaise Faugeras, Hervé Guillard, Holger Heumann, Sebastian Minjeaud, Boniface Nkonga, Richard Pasquetti, Afeintou Sangam.

The main research topics are:

1. Modelling and analysis
 - Fluid closure in plasma
 - Turbulence
 - Plasma anisotropy type instabilities
 - Free boundary equilibrium (FBE)
 - Coupling FBE – Transport
2. Numerical methods and simulations
 - High order methods
 - Curvilinear coordinate systems
 - Equilibrium simulation
 - Pressure correction scheme
 - Anisotropy
 - Solving methods and parallelism
3. Identification and control
 - Inverse problem: Equilibrium reconstruction
 - Open loop control
4. Applications
 - MHD instabilities : Edge-Localized Modes (ELMs)
 - Edge plasma turbulence
 - Optimization of scenarii

COFFEE Project-Team

3. Research Program

3.1. Research Program

Mathematical modeling and computer simulation are among the main research tools for environmental management, risks evaluation and sustainable development policy. Many aspects of the computer codes as well as the PDEs systems on which these codes are based can be considered as questionable regarding the established standards of applied mathematical modeling and numerical analysis. This is due to the intricate multiscale nature and tremendous complexity of those phenomena that require to set up new and appropriate tools. Our research group aims to contribute to bridging the gap by developing advanced abstract mathematical models as well as related computational techniques.

The scientific basis of the proposal is two-fold. On the one hand, the project is “technically-driven”: it has a strong content of mathematical analysis and design of general methodology tools. On the other hand, the project is also “application-driven”: we have identified a set of relevant problems motivated by environmental issues, which share, sometimes in a unexpected fashion, many common features. The proposal is precisely based on the conviction that these subjects can mutually cross-fertilize and that they will both be a source of general technical developments, and a relevant way to demonstrate the skills of the methods we wish to design.

To be more specific:

- We consider evolution problems describing highly heterogeneous flows (with different phases or with high density ratio). In turn, we are led to deal with non linear systems of PDEs of convection and/or convection–diffusion type.
- The nature of the coupling between the equations can be two-fold, which leads to different difficulties, both in terms of analysis and conception of numerical methods. For instance, the system can couple several equations of different types (elliptic/parabolic, parabolic/hyperbolic, parabolic or elliptic with algebraic constraints, parabolic with degenerate coefficients....). Furthermore, the unknowns can depend on different sets of variables, a typical example being the fluid/kinetic models for particulate flows. In turn, the simulation cannot use a single numerical approach to treat all the equations. Instead, hybrid methods have to be designed which raise the question of fitting them in an appropriate way, both in terms of consistency of the discretization and in terms of stability of the whole computation. For the problems under consideration, the coupling can also arise through interface conditions. It naturally occurs when the physical conditions are highly different in subdomains of the physical domain in which the flows takes place. Hence interface conditions are intended to describe the exchange (of mass, energy...) between the domains. Again it gives rise to rather unexplored mathematical questions, and for numerics it yields the question of defining a suitable matching at the discrete level, that is requested to preserve the properties of the continuous model.
- By nature the problems we wish to consider involve many different scales (of time or length basically). It raises two families of mathematical questions. In terms of numerical schemes, the multiscale feature induces the presence of stiff terms within the equations, which naturally leads to stability issues. A clear understanding of scale separation helps in designing efficient methods, based on suitable splitting techniques for instance. On the other hand asymptotic arguments can be used to derive hierarchy of models and to identify physical regimes in which a reduced set of equations can be used.

We can distinguish the following fields of expertise

- Numerical Analysis: Finite Volume Schemes, Well-Balanced and Asymptotic-Preserving Methods
 - Finite Volume Schemes for Diffusion Equations
 - Finite Volume Schemes for Conservation Laws
 - Well-Balanced and Asymptotic-Preserving Methods
- Modeling and Analysis of PDEs
 - Kinetic equations and hyperbolic systems
 - PDEs in random media
 - Interface problems

FLUMINANCE Project-Team

3. Research Program

3.1. Estimation of fluid characteristic features from images

The measurement of fluid representative features such as vector fields, potential functions or vorticity maps, enables physicists to have better understanding of experimental or geophysical fluid flows. Such measurements date back to one century and more but became an intensive subject of research since the emergence of correlation techniques [48] to track fluid movements in pairs of images of a particles laden fluid or by the way of clouds photometric pattern identification in meteorological images. In computer vision, the estimation of the projection of the apparent motion of a 3D scene onto the image plane, referred to in the literature as optical-flow, is an intensive subject of researches since the 80's and the seminal work of B. Horn and B. Schunk [61]. Unlike to dense optical flow estimators, the former approach provides techniques that supply only sparse velocity fields. These methods have demonstrated to be robust and to provide accurate measurements for flows seeded with particles. These restrictions and their inherent discrete local nature limit too much their use and prevent any evolutions of these techniques towards the devising of methods supplying physically consistent results and small scale velocity measurements. It does not authorize also the use of scalar images exploited in numerous situations to visualize flows (image showing the diffusion of a scalar such as dye, pollutant, light index refraction, fluorecein,...). At the opposite, variational techniques enable in a well-established mathematical framework to estimate spatially continuous velocity fields, which should allow more properly to go towards the measurement of smaller motion scales. As these methods are defined through PDE's systems they allow quite naturally constraints to be included such as kinematic properties or dynamic laws governing the observed fluid flows. Besides, within this framework it is also much easier to define characteristic features estimation procedures on the basis of physically grounded data model that describes the relation linking the observed luminance function and some state variables of the observed flow. The Fluminance group has allowed a substantial progress in this direction with the design of dedicated dense estimation techniques to estimate dense fluid motion fields. See [7] for a detailed review. More recently problems related to scale measurement and uncertainty estimation have been investigated [55]. Dynamically consistent and highly robust techniques have been also proposed for the recovery of surface oceanic streams from satellite images [51]. Very recently parameter-free approaches relying on uncertainty concept has been devised [52]. This technique outperforms the state of the art.

3.2. Data assimilation and Tracking of characteristic fluid features

Real flows have an extent of complexity, even in carefully controlled experimental conditions, which prevents any set of sensors from providing enough information to describe them completely. Even with the highest levels of accuracy, space-time coverage and grid refinement, there will always remain at least a lack of resolution and some missing input about the actual boundary conditions. This is obviously true for the complex flows encountered in industrial and natural conditions, but remains also an obstacle even for standard academic flows thoroughly investigated in research conditions.

This unavoidable deficiency of the experimental techniques is nevertheless more and more compensated by numerical simulations. The parallel advances in sensors, acquisition, treatment and computer efficiency allow the mixing of experimental and simulated data produced at compatible scales in space and time. The inclusion of dynamical models as constraints of the data analysis process brings a guaranty of coherency based on fundamental equations known to correctly represent the dynamics of the flow (e.g. Navier Stokes equations) [11]. Conversely, the injection of experimental data into simulations ensures some fitting of the model with reality.

To enable data and models coupling to achieve its potential, some difficulties have to be tackled. It is in particular important to outline the fact that the coupling of dynamical models and image data are far from being straightforward. The first difficulty is related to the space of the physical model. As a matter of fact, physical models describe generally the phenomenon evolution in a 3D Cartesian space whereas images provides generally only 2D tomographic views or projections of the 3D space on the 2D image plane. Furthermore, these views are sometimes incomplete because of partial occlusions and the relations between the model state variables and the image intensity function are otherwise often intricate and only partially known. Besides, the dynamical model and the image data may be related to spatio-temporal scale spaces of very different natures which increases the complexity of an eventual multiscale coupling. As a consequence of these difficulties, it is necessary generally to define simpler dynamical models in order to assimilate image data. This redefinition can be done for instance on an uncertainty analysis basis, through physical considerations or by the way of data based empirical specifications. Such modeling comes to define inexact evolution laws and leads to the handling of stochastic dynamical models. The necessity to make use and define sound approximate models, the dimension of the state variables of interest and the complex relations linking the state variables and the intensity function, together with the potential applications described earlier constitute very stimulating issues for the design of efficient data-model coupling techniques based on image sequences.

On top of the problems mentioned above, the models exploited in assimilation techniques often suffer from some uncertainties on the parameters which define them. Hence, a new emerging field of research focuses on the characterization of the set of achievable solutions as a function of these uncertainties. This sort of characterization indeed turns out to be crucial for the relevant analysis of any simulation outputs or the correct interpretation of operational forecasting schemes. In this context, stochastic modeling play a crucial role to model and process uncertainty evolution along time. As a consequence, stochastic parameterization of flow dynamics has already been present in many contributions of the Fluminance group in the last years and will remain a cornerstone of the new methodologies investigated by the team in the domain of uncertainty characterization.

This wide theme of research problems is a central topic in our research group. As a matter of fact, such a coupling may rely on adequate instantaneous motion descriptors extracted with the help of the techniques studied in the first research axis of the FLUMINANCE group. In the same time, this coupling is also essential with respect to visual flow control studies explored in the third theme. The coupling between a dynamics and data, designated in the literature as a Data Assimilation issue, can be either conducted with optimal control techniques [62], [63] or through stochastic filtering approaches [56], [59]. These two frameworks have their own advantages and deficiencies. We rely indifferently on both approaches.

3.3. Optimization and control of fluid flows with visual servoing

Fluid flow control is a recent and active research domain. A significant part of the work carried out so far in that field has been dedicated to the control of the transition from laminarity to turbulence. Delaying, accelerating or modifying this transition is of great economical interest for industrial applications. For instance, it has been shown that for an aircraft, a drag reduction can be obtained while enhancing the lift, leading consequently to limit fuel consumption. In contrast, in other application domains such as industrial chemistry, turbulence phenomena are encouraged to improve heat exchange, increase the mixing of chemical components and enhance chemical reactions. Similarly, in military and civilians applications where combustion is involved, the control of mixing by means of turbulence handling rouses a great interest, for example to limit infra-red signatures of fighter aircraft.

Flow control can be achieved in two different ways: passive or active control. Passive control provides a permanent action on a system. Most often it consists in optimizing shapes or in choosing suitable surfacing (see for example [54] where longitudinal riblets are used to reduce the drag caused by turbulence). The main problem with such an approach is that the control is, of course, inoperative when the system changes. Conversely, in active control the action is time varying and adapted to the current system's state. This approach requires an external energy to act on the system through actuators enabling a forcing on the flow through for instance blowing and suction actions [66], [58]. A closed-loop problem can be formulated as an optimal control

issue where a control law minimizing an objective cost function (minimization of the drag, minimization of the actuators power, etc.) must be applied to the actuators [50]. Most of the works of the literature indeed comes back to open-loop control approaches [65], [60], [64] or to forcing approaches [57] with control laws acting without any feedback information on the flow actual state. In order for these methods to be operative, the model used to derive the control law must describe as accurately as possible the flow and all the eventual perturbations of the surrounding environment, which is very unlikely in real situations. In addition, as such approaches rely on a perfect model, a high computational costs is usually required. This inescapable pitfall has motivated a strong interest on model reduction. Their key advantage being that they can be specified empirically from the data and represent quite accurately, with only few modes, complex flows' dynamics. This motivates an important research axis in the Fluminance group.

3.4. Numerical models applied to hydrogeology and geophysics

The team is strongly involved in numerical models for hydrogeology and geophysics. There are many scientific challenges in the area of groundwater simulations. This interdisciplinary research is very fruitful with cross-fertilizing subjects.

In geophysics, a main concern is to solve inverse problems in order to fit the measured data with the model. Generally, this amounts to solve a linear or nonlinear least-squares problem.

Models of geophysics are in general coupled and multi-physics. For example, reactive transport couples advection-diffusion with chemistry. Here, the mathematical model is a set of nonlinear Partial Differential Algebraic Equations. At each timestep of an implicit scheme, a large nonlinear system of equations arise. The challenge is to solve efficiently and accurately these large nonlinear systems.

3.5. Numerical algorithms and high performance computing

Linear algebra is at the kernel of most scientific applications, in particular in physical or chemical engineering. The objectives are to analyze the complexity of these different methods, to accelerate convergence of iterative methods, to measure and improve the efficiency on parallel architectures, to define criteria of choice.

LEMON Team

3. Research Program

3.1. Inland flow processes

3.1.1. *Shallow water models with porosity*

State of the Art

Simulating urban floods and free surface flows in wetlands requires considerable computational power. Two-dimensional shallow water models are needed. Capturing the relevant hydraulic detail often requires computational cell sizes smaller than one meter. For instance, meshing a complete urban area with a sufficient accuracy would require 10^6 to 10^8 cells, and simulating one second often requires several CPU seconds. This makes the use of such model for crisis management impossible. Similar issues arise when modelling wetlands and coastal lagoons, where large areas are often connected by an overwhelming number of narrow channels, obstructed by vegetation and a strongly variable bathymetry. Describing such channels with the level of detail required in a 2D model is impracticable. A new generation of models overcoming this issue has emerged over the last 20 years: porosity-based shallow water models. They are obtained by averaging the two-dimensional shallow water equations over large areas containing both water and a solid phase [35]. The size of a computational cell can be increased by a factor 10 to 50 compared to a 2D shallow water model, with CPU times reduced by 2 to 3 orders of magnitude [53]. While the research on porosity-based shallow water models has accelerated over the past decade [48], [64], [68], [46], [45], [53], [78], [79], [74], [47], a number of research issues remain pending.

Four year research objectives

The research objectives are (i) to improve the upscaling of the flux and source term models to be embedded in porosity shallow water models, (ii) to validate these models against laboratory and in situ measurements. Improving the upscaled flux and source term models for urban applications requires that description of anisotropy in porosity models be improved to account for the preferential flows induced by building and street alignment. The description of the porosity embedded in the most widespread porosity approach, the so-called Integral Porosity model [64] [5], has been shown to provide an incomplete description of the connectivity properties of the urban medium. Firstly, the governing equations are strongly mesh-dependent because of consistency issues [5]. Secondly, the flux and source term models fail to reproduce the alignment with the main street axes in a number of situations [4]. Another path for improvement concerns the upscaling of obstacle-induced drag terms in the presence of complex geometries. Recent upscaling research results (to be submitted for publication within the next few weeks) obtained by the LEMON team in collaboration with Tour du Valat suggest that the effects of microtopography on the flow cannot be upscaled using "classical" equation-of-state approaches, as done in most hydraulic models. A totally different approach must be proposed. The next four years will be devoted to the development and validation of improved flux and source term closures in the presence of strongly anisotropic urban geometries and in the presence of strongly variable topography. Validation will involve not only the comparison of porosity model outputs with refined flow simulation results, but also the validation against experimental data sets. No experimental data set allowing for a sound validation of flux closures in porosity models can be found in the literature. Laboratory experiments will be developed specifically in view of the validation of porosity models. Such experiments will be set up and carried out in collaboration with the Université Catholique de Louvain (UCL), that has an excellent track record in experimental hydraulics and the development of flow monitoring and data acquisition equipment. These activities will take place in the framework of the PoroCity Associate International Laboratory (see next paragraph).

People

Carole DELENNE , Vincent GUINOT , Antoine ROUSSEAU

External collaborations

- Tour du Valat (O. Boutron): the partnership with TdV focuses on the development and application of depth-dependent porosity models to the simulation of coastal lagoons, where the bathymetry and geometry is too complex to be represented using refined flow models.
- University of California Irvine (B. Sanders): the collaboration with UCI started in 2014 with research on the representation of urban anisotropic features in integral porosity models [53]. It has led to the development of the Dual Integral Porosity model [6]. Ongoing research focuses on improved representations of urban anisotropy in urban floods modelling.
- Université Catholique de Louvain - UCL (S. Soares-Frazão): UCL is one of the few places with experimental facilities allowing for the systematic, detailed validation of porosity models. The collaboration with UCL started in 2005 and will continue with the PoroCity Associate International Laboratory proposal. In this proposal, a for year research program is set up for the validation, development and parametrization of shallow water models with porosity.

3.1.2. Forcing

State of the Art

Reproducing optimally realistic spatio-temporal rainfall fields is of salient importance to the forcing of hydrodynamic models. This challenging task requires combining intense, usual and dry weather events. Far from being straightforward, this combination of extreme and non-extreme scenarios requires a realistic modelling of the transitions between normal and extreme periods. [57] have proposed in a univariate framework a statistical model that can serve as a generator and that takes into account low, moderate and intense precipitation. In the same vein, [75] developed a bivariate model. However, its extension to a spatial framework remains a challenge. Existing spatial precipitation stochastic generators are generally based on Gaussian spatial processes [22], [55], that are not adapted to generate extreme rainfall events. Recent advances in spatio-temporal extremes modelling based on generalized Pareto processes [39], [71] and semi-parametric simulation techniques [28] are very promising and could form the base for relevant developments in our framework.

Four year research objectives

The purpose is to develop stochastic methods for the simulation of realistic spatio-temporal processes integrating extreme events. Two steps are identified. The first one is about the simulation of extreme events and the second one concerns the combination of extreme and non extreme events in order to build complete, realistic precipitations time series. As far as the first step is concerned, a first task will be to understand and to model the space-time structure of hydrological extremes such as those observed in the French Mediterranean basin, that is known for its intense rainfall events (Cevenol episodes), which have recently received increased attention. We will propose modelling approaches based on the exceedance, which allows the simulated fields to be interpreted as events. Parametric, semi-parametric and non-parametric approaches are currently under consideration. They would allow a number of scientific locks to be removed. Examples of such locks are e.g. accounting for the temporal dimension and for various dependence structures (asymptotic dependence or asymptotic independence possibly depending on the dimension and/or the distance considered). Methodological aspects are detailed in Section 3.3.1. The second step, that is not straightforward, consists in combining different spatio-temporal simulations in order to help to ultimately develop a stochastic precipitation generator capable of producing full precipitation fields, including dry and non-extreme wet periods.

People

Carole DELENNE , Vincent GUINOT , Gwladys TOULEMONDE

External collaborations

The Cerise (2016-2018) project, led by Gwladys TOULEMONDE , is funded by INSU via the action MANU (MAthematical and Numerical methods) of the LEFE program. It aims to propose methods for simulating scenarii integrating spatio-temporal extremes fields with a possible asymptotic independence for impact studies in environmental sciences. Among the members of this project, Jean-Noel Bacro (IMAG, UM), Carlo Gaetan (DAIS, Italy) and Thomas Opitz (BioSP, MIA, INRA) are involved in the first step as identified in the research objectives of the present sub-section. Denis Allard (BioSP, MIA, INRA) and Julie Carreau (IRD, HSM) will be involved in the second one.

3.1.3. Inland hydrobiological systems**State of the Art**

Water bodies such as lakes or coastal lagoons (possibly connected to the sea) located in high human activity areas are subject to various kinds of stress such as industrial pollution, high water demand or bacterial blooms caused by freshwater over-enrichment. For obvious environmental reasons, these water resources have to be protected, hence the need to better understand and possibly control such fragile ecosystems to eventually develop decision-making tools. From a modelling point of view, they share a common feature in that they all involve interacting biological and hydrological processes. According to [40], models may be classified into two main types: “minimal dynamic models” and “complex dynamic models”. These two model types do not have the same objectives. While the former are more heuristic and rather depict the likelihood of considered processes, the latter are usually derived from fundamental laws of biochemistry or fluid dynamics. Of course, the latter necessitate much more computational resources than the former. In addition, controlling such complex systems (usually governed by PDEs) is by far more difficult than controlling the simpler ODE-driven command systems.

LEMON has already contributed both to the reduction of PDE models for the simulation of water confinement in coastal lagoons [41], [23] and to the improvement of ODE models in order to account for space-heterogeneity of bioremediation processes in water resources [21].

Four year research objectives

In collaboration with colleagues from the ANR-ANSWER project and colleagues from INRA, our ambition is to improve existing models of lagoon/marine ecosystems by integrating both accurate and numerically affordable coupled hydrobiological systems. A major challenge is to find an optimal trade-off between the level of detail in the description of the ecosystem and the level of complexity in terms of number of parameters (in particular regarding the governing equations for inter-species reactions). The model(s) should be able to reproduce the inter-annual variability of the observed dynamics of the ecosystem in response to meteorological forcing. This will require the adaptation of hydrodynamics equations to such time scales (reduced/upscaled models such as porosity shallow water models (see Section 3.1.1) will have to be considered) together with the coupling with the ecological models. At short time scales (i.e. the weekly time scale), accurate (but possibly CPU-consuming) 3D hydrodynamic models processes (describing thermal stratification, mixing, current velocity, sediment resuspension, wind waves...) are needed. On the longer term, it is intended to develop reduced models accounting for spatial heterogeneity.

The team will focus on two main application projects in the coming years:

- the ANR ANSWER project (2017-2021, with INRA Montpellier and LEESU) focusing on the cyanobacteria dynamics in lagoons and lakes. A PhD student will be co-advised by Antoine ROUSSEAU in collaboration with Céline Casenave (INRA, Montpellier).
- the long term collaboration with Alain Rapaport (INRA Montpellier) will continue both on the bioremediation of water resources such as the Tunquen lagoon in Chile and with a new ongoing project on water reuse (converting wastewater into water that can be reused for other purposes such as irrigation of agricultural fields). Several projects are submitted to the ANR and local funding structures in Montpellier.

People

Antoine ROUSSEAU , Vincent GUINOT , Joseph Kahn, PhD student (march 2018)

External collaborations

- ANR ANSWER consortium: Céline Casenave (UMR MISTEA, INRA Montpellier), Brigitte Vinçon-Leite (UM LEESU, ENPC), Jean-François Humbert (UMR IEES, UPMC). ANSWER is a French-Chinese collaborative project that focuses on the modelling and simulation of eutrophic lake ecosystems to study the impact of anthropogenic environmental changes on the proliferation of cyanobacteria. Worldwide the current environmental situation is preoccupying: man-driven water needs increase, while the quality of the available resources is deteriorating due to pollution of various kinds and to hydric stress. In particular, the eutrophication of lentic ecosystems due to excessive inputs of nutrients (phosphorus and nitrogen) has become a major problem because it promotes cyanobacteria blooms, which disrupt the functioning and the uses of the ecosystems.
- A. Rousseau has a long lasting collaboration with Alain Rapaport (UMR MISTEA, INRA Montpellier) and Héctor Ramirez (CMM, Université du Chili).

3.1.4. Parametrization

State of the Art

Numerical modelling requires data acquisition, both for model validation and for parameter assessment. Model benchmarking against laboratory experiments is an essential step and is essential to team's strategy. However, scale model experiments may have several drawbacks: *i*) experiments are very expensive and extremely time-consuming, *ii*) experiments cannot always be replicated, and measurement have precision and reliability limitations, *iii*) dimensional similarity (in terms of geometry and flow characteristic variables such as Froude or Reynolds numbers) cannot always be preserved.

An ideal way to obtain data would be to carry out in situ measurements. But this would be too costly at the scale of studied systems, not to mention the fact that field may become impracticable during flood periods.

Remote sensing data are becoming widely available with high spatial and temporal resolutions. Several recent studies have shown that flood extends can be extracted from optical or radar images [42], for example: to characterize the flood dynamics of great rivers [58], to monitor temporary ponds [69], but also to calibrate hydrodynamics models and assess roughness parameters [66], [49], [77].

Upscaled models developed in LEMON embed new parameters that reflect the statistical properties of the medium geometry. Two types of information are needed: the directional properties of the medium and its flow connectivity properties. New methods are thus to be developed to characterize such statistical properties from geographical data.

Four year research objectives

This research line consists in deriving methods and algorithms for the determination of upscaled model parameters from various types and sources of geodata: aerial photographs, urban databases, remote sensing SAR or optical data, etc. In developed countries, it is intended to extract information on the porosities and their principal directions from urban or National geographical survey databases. Such databases usually incorporate separate layers for roads, buildings, parking lots, yards, etc. Most of the information is stored in vector form, which can be expected to make the treatment of urban anisotropic properties easier than with the raster format. Moreover, data is made increasingly available over the world thanks to crowdsourcing (e.g. OpenStreetMap). However, in order to achieve a correct parametrization of a porosity model, identifying areas with homogeneous porosity properties is necessary. Algorithms identifying the shape and extension of such areas are still to be developed.

In developing countries, such level of detail in vector format may not be available. Moreover, vector data for the street network does not provide all the relevant information. In suburban areas, lawns, parks and other vegetated areas may also contribute to flood propagation and storage. In this context, it is intended to extract the necessary information from aerial and/or satellite images, that are widely available and the spatial resolution of which improves constantly. A major research line will consist in deriving the information on street preferential orientation using textural analysis techniques. Such techniques have been used successfully in the field of agricultural pattern identification during Carole DELENNE PhD thesis [37], [62]. However, their application to the urban environment raises a number of issues. One of them is the strongly discontinuous character of the urban medium, that makes textural analysis difficult.

In wetlands applications, the flow connectivity is a function of the free surface elevation. Characterizing such connectivity requires that topographical variations be known with high accuracy. Despite the increased availability of direct topographic measurements from LiDARS on riverine systems, data collection remains costly when wide areas are involved. Data acquisition may also be difficult when poorly accessible areas are dealt with. If the amount of topographic points is limited, information on elevation contour lines can be easily extracted from the flood dynamics visible in simple SAR or optical images. A challenge is thus to use such data in order to estimate continuous topography on the floodplain combining heterogeneous data: topographic sampling points and located contour lines the levels of which are unknown or uncertain.

People

Carole DELENNE , Vincent GUINOT , Antoine ROUSSEAU

External collaborations

- The methodologies concerning geographical databases in vector form will be developed in strong collaboration with C. Dieulin at HSM in the framework of the PoroCity Associate International Laboratory cited above.
- Research on topography reconstruction in wetlands begun in collaboration with J.-S. Bailly (LISAH) in 2016 [36] and will continue in the coming years.

3.2. Marine and coastal systems

3.2.1. Multi-scale ocean modelling

State of the Art

In physical oceanography, all operational models - regardless of the scale they apply to - are derived from the complete equations of geophysical fluid dynamics. Depending on the considered process properties (nonlinearity, scale) and the available computational power, the original equations are adapted with some simplifying hypotheses. The reader can refer to [63], [56] for a hierarchical presentation of such models.

In the nearshore area, the hydrostatic approximation that is used in most large scale models (high sea) cannot be used without a massive loss of accuracy. In particular, shallow water models are inappropriate to describe the physical processes that occur in this zone (see Figure 1). This is why Boussinesq-type models are preferred: see [54]. They embed dispersive terms that allow for shoaling and other bathymetry effects. Since the pioneering works of Green and Naghdi (see [43]), numerous theoretical and numerical studies have been delivered by the "mathematical oceanography" community, more specifically in France (see the works of Lannes, Marche, Sainte-Marie, Bresch, etc.). The corresponding numerical models (BOSZ, WaveBox) must thus be integrated in any reasonable nearshore modelling platform.

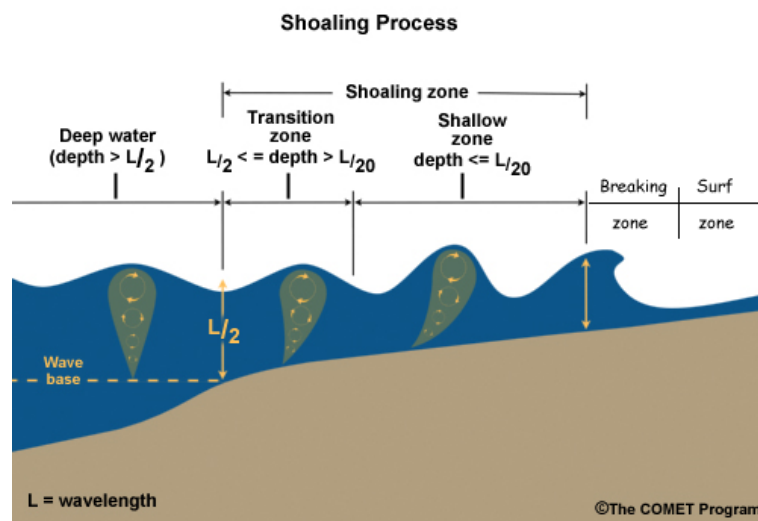


Figure 1. Deep sea, shoaling, breaking and surf zones.

However, these models cannot simply replace all previous models everywhere in the ocean: dispersive models are useless away from the shore and it is known that wave breaking cannot be simulated using Boussinesq-type equations. Hence the need to couple these models with others. Some work has been done in this direction with a multi-level nesting using software packages such as ROMS, but to the best of our knowledge, all the "boxes" rely on the same governing equations with different grid resolutions. A real coupling between different models is a more difficult task since different models may have different mathematical properties, as shown in the work by Blayo and Rousseau on shallow water modelling (see [1] and [24]).

Four year research objectives

Starting from the knowledge acquired in the collaboration with Eric Blayo on model coupling using domain decomposition techniques, our ambition is to propose theoretical and numerical tools in order to incorporate nearshore ocean models into large complex systems including several space and time scales. Two complementary research directions are considered:

- **Dispersive vs non-dispersive shallow water models.** As depicted in Figure 1 above, Boussinesq-type models (embedding dispersive effects) should be used in the so-called shoaling zone. The coupling with classical deep-sea / shallow water models has to be done such that all the processes in Figure 1 are correctly modelled (by different equations), with a reduced numerical cost. As a first guess, we think that Schwarz-type methods (widely used by the DDM community) could be good candidates, in particular when the interface locations are well-known. Moving interfaces (depending

on the flow, the bathymetry and naturally the wind and all external forcings) is a more challenging objective that will be tackled after the first step (known interface) is achieved.

- **spectral vs time-domain models.** In the context of mathematical modelling and numerical simulation for the marine energy, we want to build a coupled numerical model that would be able to simulate wave propagation in domains covering both off-shore regions, where spectral models are used, and nearshore regions, better described by nonlinear dispersive (Boussinesq-type) models. While spectral models work with a statistical and phase-averaged description of the waves, solving the evolution of its energy spectrum, Boussinesq-type models are phase-resolving and solves nonlinear dispersive shallow water equations for physical variables (surface elevation and velocity) in the time domain. Furthermore, the time and space scales are very different: they are much larger in the case of spectral models, which justifies their use for modelling off-shore propagation over large time frames. Moreover, important small scale phenomena in nearshore areas are better captured by Boussinesq models, in which the time step is limited by the CFL condition. From a mathematical and modelling point of view, this task mainly consists in working on the boundary conditions of each model, managing the simultaneous use of spectral and time series data, while studying transparent boundary conditions for the models and developing domain decomposition approaches to improve the exchange of information.

People

Antoine ROUSSEAU , Joao Caldas

External collaborations

- **Eric Blayo** is the former scientific leader of team MOISE in Grenoble, where Antoine ROUSSEAU was first recruited. Blayo and Rousseau have co-advised 3 PhDs and continue to work together on coupling methods in hydrodynamics, especially in the framework of the **COMODO** ANR network.
- **Fabien Marche** (at IMAG, Montpellier, currently on leave in Bordeaux) is an expert in numerical modelling and analysis of Boussinesq-type models. He is the principal investigator of the WaveBox software project, to be embedded in the national scale Uhaina initiative.
- In the framework of its collaboration with **MERIC**, Antoine ROUSSEAU and Joao collaborate with the consortium DiMe (ANR-FEM project), and more particularly with Jean-François Filipot and Volker Roeber for the coupling of spectral and time-domain methods.

3.2.2. Data-model interactions

State of the Art

An alternative to direct observations is the chaining of numerical models, which for instance represent the physic from offshore to coastal areas. Typically, output data from atmospheric and ocean circulation models are used as forcings for a wave model, which in turn feeds a littoral model. In the case of extreme events, their numerical simulation from physical models is generally unreachable. This is due to a lack of knowledge on boundary conditions and on their physical reliability for such extreme quantities. Based on numerical simulated data, an alternative is to use statistical approaches. [28] proposed such an approach. They first produced and studied a 52-year hindcast using the WW3 wave model [26], [29], [27], [72]. Then stemming from parts of the original work of [25], [44], [39], [28] proposed a semi-parametric approach which aims to simulate extreme space-time waves processes to, in turn, force a littoral hazard model. Nevertheless their approach allows only a very small number of scenarios to be simulated.

Four year research objectives

A first objective is to establish the link between the simulation approach proposed by [28] and the Pareto Processes [39]. This will allow the work of [28] to be generalized, thus opening up the possibility of generating an infinity of extreme scenarios. While continuing to favor the semi- or non-parametric approaches made possible by the access to high spatial resolution calculations, we will try to capture the strength of potentially decreasing extremal dependence when moving towards higher values, which requires the development of models that allow for so-called asymptotic independence.

People

Antoine ROUSSEAU , Gwladys TOULEMONDE , Fátima Palacios Rodríguez

External collaborations

- The collaboration with Romain Chailan (TwinSolutions, Montpellier) and Frédéric Bouchette (Geosciences, UM) started in 2012 during the PhD of Romain entitled Application of scientific computing and statistical analysis to address coastal hazards.
- During her post doctoral position, Fátima Palacios Rodríguez with her co-advisors will consider a generalization of the proposed simulation method by [28].

3.3. Methodological developments

In addition to the application-driven sections, the team will also work on the following theoretical questions. They are clearly connected to the abovementioned scientific issues but do not correspond to a specific application or process.

3.3.1. Stochastic models for extreme events

State of the Art

Max-stable random fields [67], [65], [52], [32], [59] are the natural limit models for spatial maximum data and have spawned a very rich literature. An overview of typical approaches to modelling maxima is due to [34]. Physical interpretation of simulated data from such models can be discussed. An alternative to the max-stable framework are models for threshold exceedances. Processes called GPD processes, which appears as a generalization of the univariate formalism of the high thresholds exceeding a threshold based on the GPD, have been proposed [39], [71]. Strong advantages of these thresholding techniques are their capability to exploit more information from the data and explicitly model the original event data. However, the asymptotic dependence stability in these limiting processes for maximum and threshold exceedance tends to be overly restrictive when asymptotic dependence strength decreases at high levels and may ultimately vanish in the case of asymptotic independence. Such behaviours appear to be characteristic for many real-world data sets such as precipitation fields [33], [70]. This has motivated the development of more flexible dependence models such as max-mixtures of max-stable and asymptotically independent processes [76], [20] for maxima data, and Gaussian scale mixture processes [60], [51] for threshold exceedances. These models can accommodate asymptotic dependence, asymptotic independence and Gaussian dependence with a smooth transition. Extreme events also generally present a temporal dependence [73]. Developing flexible space-time models for extremes is crucial for characterizing the temporal persistence of extreme events spanning several time steps; such models are important for short-term prediction in applications such as the forecasting of wind power and for extreme event scenario generators providing inputs to impact models, for instance in hydrology and agriculture. Currently, only few models are available from the statistical literature (see for instance [30], [31], [50]) and remain difficult to interpret.

Four year research objectives

The objective is to extend state-of-the-art methodology with respect to three important aspects: 1) adapting well-studied spatial modelling techniques for extreme events based on asymptotically justified models for threshold exceedances to the space-time setup; 2) replacing restrictive parametric dependence modelling by semiparametric or nonparametric approaches; 3) proposing more flexible spatial models in terms of asymmetry or in terms of dependence. This means being able to capture the strength of potentially decreasing extremal dependence when moving towards higher values, which requires developing models that allow for so-called asymptotic independence.

People

Gwladys TOULEMONDE , Fátima Palacios Rodríguez

External collaborations

In a natural way, the Cerise project members are the main collaborators for developing and studying new stochastic models for extremes.

- More specifically, research with Jean-Noel Bacro (IMAG, UM), Carlo Gaetan (DAIS, Italy) and Thomas Opitz (BioSP, MIA, INRA) focuses on relaxing dependence hypothesis.
- The asymmetry issue and generalization of some Copula-based models are studied with Julie Carreau (IRD, HydroSciences, UM).

3.3.2. Integrating heterogeneous data

State of the Art

Assuming that a given hydrodynamic models is deemed to perform satisfactorily, this is far from being sufficient for its practical application. Accurate information is required concerning the overall geometry of the area under study and model parametrization is a necessary step towards the operational use. When large areas are considered, data acquisition may turn out prohibitive in terms of cost and time, not to mention the fact that information is sometimes not accessible directly on the field. To give but one example, how can the roughness of an underground sewer pipe be measured?

A strategy should be established to benefit from all the possible sources of information in order to gather data into a geographical database, along with confidence indexes.

Four year research objectives

The assumption is made that even hardly accessible information often exists. This stems from the increasing availability of remote-sensing data, to the crowdsourcing of geographical databases, including the inexhaustible source of information provided by the internet. However, information remains quite fragmented and stored in various formats: images, vector shapes, texts, etc. The overall objective of this research line is to develop methodologies to gather various types of data in the aim of producing an accurate mapping of the studied systems for hydrodynamics models.

One possible application concerns the combination of on-field elevation measurements and level lines extracted from remote sensing data in the aim of retrieving the general topography of the domain (see section 3.1.4).

Another application studied in the Cart'Eaux project, is the production of regular and complete mapping of urban sewer systems. Contrary to drinkable water networks, the knowledge of sewer pipe location is not straightforward, even in developed countries. The methodology applied consists in inferring the shape of the network from a partial dataset of manhole covers that can be detected from aerial images [61]. Since manhole covers positions are expected to be known with low accuracy (positional uncertainty, detection errors), a stochastic algorithm will be set up to provide a set of probable network geometries. As more information is required for hydraulic modelling than the simple mapping of the network (slopes, diameters, materials, etc.), text data mining is used to extract characteristics from data posted on the Web or available through governmental or specific databases. Using an appropriate keyword list, the web is scanned for text documents. Thematic entities are identified and linked to the surrounding spatial and temporal entities in order to ease the burden of data collection. It is clear at this stage that obtaining numerical values on specific pipes will be challenging. Thus, when no information is found, decision rules will be used to assign acceptable numerical values to enable the final hydraulic modelling.

In any case, the confidence associated to each piece of data, be it directly measured or reached from a roundabout route, should be assessed and taken into account in the modelling process. This can be done by generating a set of probable inputs (geometry, boundary conditions, forcing, etc.) yielding simulation results along with the associated uncertainty.

People

Carole DELENNE , Vincent GUINOT , Gwladys TOULEMONDE , Benjamin Commandré

External collaborations

- The Cart'Eaux project has been a lever to develop a collaboration with Berger-Levrault company and several multidisciplinary collaborations for image treatment (LIRMM), text analysis (LIRMM and TETIS) and network cartography (LISAH, IFSTTAR). These collaborations will continue with the submission of new projects.
- The problematic of inferring a connected network from scarce or uncertain data is common to several research topics in LEMON such as sewage or drainage systems, urban media and wetlands. A generic methodology will be developed in collaboration with J.-S. Bailly (LISAH).

3.3.3. Numerical methods for porosity models

State of the Art

Porosity-based shallow water models are governed by hyperbolic systems of conservation laws. The most widespread method used to solve such systems is the finite volume approach. The fluxes are computed by solving Riemann problems at the cell interfaces. This requires that the wave propagation properties stemming from the governing equations be known with sufficient accuracy. Most porosity models, however, are governed by non-standard hyperbolic systems.

Firstly, the most recently developed DIP models include a momentum source term involving the divergence of the momentum fluxes [6]. This source term is not active in all situations but takes effect only when positive waves are involved [46] [4]. The consequence is a discontinuous flux tensor and discontinuous wave propagation properties. The consequences of this on the existence and uniqueness of solutions to initial value problems (especially the Riemann problem) are not known. Nor are the consequences on the accuracy of the numerical methods used to solve this new type of equations.

Secondly, most applications of these models involve anisotropic porosity fields [53], [64]. Such anisotropy can be modelled using 2×2 porosity tensors, with principal directions that are not aligned with those of the Riemann problems in two dimensions of space. The solution of such Riemann problems has not been investigated yet. Moreover, the governing equations not being invariant by rotation, their solution on unstructured grids is not straightforward.

Thirdly, the Riemann-based, finite volume solution of the governing equations require that the Riemann problem be solved in the presence of a porosity discontinuity. While recent work [38] has addressed the issue for the single porosity equations, similar work remains to be done for integral- and multiple porosity-based models.

Four year research objectives

The four year research objectives are the following:

- investigate the properties of the analytical solutions of the Riemann problem for a continuous, anisotropic porosity field,
- extend the properties of such analytical solutions to discontinuous porosity fields,
- derive accurate and CPU-efficient approximate Riemann solvers for the solution of the conservation form of the porosity equations.

People

Vincent GUINOT , Antoine ROUSSEAU

External collaborations

Owing to the limited staff of the Lemon team, external collaborations will be sought with researchers in applied mathematics. Examples of researchers working in the field are

- Minh Le, Saint Venant laboratory, Chatou (France)
- M.E. Vazquez-Cendon, Univ. Santiago da Compostela (Spain)
- A. Ferrari, R. Vacondio, S. Dazzi, P. Mignosa, Univ. Parma (Italy)
- O. Delestre, Univ. Nice-Sophia Antipolis (France)
- F. Benkhaldoun, Univ. Paris 13 (France)

MAGIQUE-3D Project-Team

3. Research Program

3.1. Introduction

Probing the invisible is a quest that is shared by a wide variety of scientists such as archaeologists, geologists, astrophysicists, physicists, etc... Magique-3D is involved in Geophysical imaging which aims at understanding the internal structure of the Earth from the propagation of waves. Both qualitative and quantitative information are required and two geophysical techniques can be used: **seismic reflection** and **seismic inversion**. Seismic reflection provides a qualitative description of the subsurface from reflected seismic waves by indicating the position of the reflectors while seismic inversion transforms seismic reflection data into a quantitative description of the subsurface. Both techniques are inverse problems based upon the numerical solution of wave equations. Oil and Gas explorations have been pioneering application domains for seismic reflection and inversion and even if numerical seismic imaging is computationally intensive, oil companies promote the use of numerical simulations to provide synthetic maps of the subsurface. This is due to the tremendous progresses of scientific computing which have pushed the limits of existing numerical methods and it is now conceivable to tackle realistic 3D problems. However, mathematical wave modeling has to be well-adapted to the region of interest and the numerical schemes which are employed to solve wave equations have to be both accurate and scalable enough to take full advantage of parallel computing. Today, geophysical imaging tackles more and more realistic problems and we can contribute to this task by improving the modeling and by deriving advanced numerical methods for solving wave problems.

Magique-3D proposes to organize its research around three main axes:

1. Mathematical modeling of multi-physics involving wave equations;
2. Supercomputing for Helmholtz problems;
3. Construction of high-order hybrid schemes.

These three research fields will be developed with the main objective of solving inverse problems dedicated to geophysical imaging.

3.2. Mathematical modeling of multi-physics involving wave equations

Wave propagation modeling is of great interest for many applications like oil and gas exploration, non destructive testing, medical imaging, etc. It involves equations which can be solved in time or frequency domain and their numerical approximation is not easy to handle, in particular when dealing with real-world problems. In both cases, the propagation domain is either infinite or its dimensions are much greater than the characteristic wavelength of the phenomenon of interest. But since wave problems are hyperbolic, the physical phenomenon can be accurately described by computing solutions in a bounded domain including the sources which have generated the waves. Until now, we have mainly worked on imaging techniques based on acoustic or elastic waves and we have developed advanced finite element software packages which are used by Total for oil exploration. Nevertheless, research on modeling must go on because there are simulations which can still not be performed because their computational cost is much too high. This is particularly true for complex tectonics involving coupled wave equations. We then propose to address the issue of coupling wave equations problems by working on the mathematical construction of reduced systems. By this way, we hope to improve simulations of elasto-acoustic and electro-seismic phenomena and then, to perform numerical imaging of strongly heterogeneous media. Even in the simplest situation where the wavelengths are similar (elasto-acoustic coupling), the dimension of the discrete coupled problem is huge and it is a genuine issue in the prospect of solving 3D inverse problems.

The accurate numerical simulation of full wave problems in heterogeneous media is computationally intensive since it needs numerical schemes based on grids. The size of the cells depends on the propagation velocity of waves. When coupling wave problems, conversion phenomena may occur and waves with very different propagation velocity coexist. The size of the cells is then defined from the smallest velocity and in most of the real-world cases, the computational cost is crippling. Regarding existing computing capabilities, we propose to derive intermediate models which require less computational burden and provide accurate solutions for a wide-ranging class of problems including Elasto-acoustics and Electro-seismology.

When it comes to mathematical analysis, we have identified two tasks which could help us simulate realistic 3D multi-physics wave problems and which are in the scope of our *savoir-faire*. They are construction of approximate and multiscale models which are different tasks. The construction of approximate problems aims at deriving systems of equations which discrete formulation involves middle-sized matrices and in general, they are based on high frequency hypothesis. Multiscale models are based on a rigorous analysis involving a small parameter which does not depend on the propagation velocity necessarily.

Recently, we have conducted research on the construction of approximate models for offshore imaging. Elastic and acoustic wave equations are coupled and we investigate the idea of eliminating the computations inside water by introducing equivalent interface conditions on the sea bottom. We apply an On-Surface-Radiation-Condition (OSRC) which is obtained from the approximation of the acoustic Dirichlet-to-Neumann (DtN) operator [74], [53]. To the best of our knowledge, OSRC method has never been used for solving reduced coupling wave problems and preliminary promising results are available at [56]. We would like to investigate this technique further because we could form a battery of problems which can be solved quickly. This would provide a set of solutions which we could use as initial guess for solving inverse problems. But we are concerned with the performance of the OSRC method when wave conversions with different wavelengths occur. Anyway, the approximation of the DtN operator is not obvious when the medium is strongly heterogeneous and multiscale analysis might be more adapted. For instance, according to existing results in Acoustics and Electromagnetism for the modeling of wire antennas [65], multiscale analysis should turn out to be very efficient when the propagation medium includes well logs, fractures and faults which are very thin structures when compared to the wavelength of seismic waves. Moreover, multiscale analysis should perform well when the medium is strongly oscillating like porous media. It could thus provide an alternative to homogenization techniques which can be applied only when the medium is periodic. We thus propose to develop reduced multi-scale models by performing rigorous mathematical procedure based on regular and singular multiscale analysis. Our approach distinguishes itself from others because it focuses on the numerical representation of small structures by time-dependent problems. This could give rise to the development of new finite element methods which would combine DG approximations with XFEM (Extended Finite Element Method) which has been created for the finite element treatment of thin structures like cracks.

But Earth imaging must be more than using elasto-acoustic wave propagation. Electromagnetic waves can also be used and in collaboration with Prof. D. Pardo (Iker Basque Foundation and University of Bilbao), we conduct researches on passive imaging to probe boreholes. Passive imaging is a recent technique of imaging which uses natural electromagnetic fields as sources. These fields are generated by hydromagnetic waves propagating in the magnetosphere which transform into electromagnetic waves when they reach the ionosphere. This is a mid-frequency imaging technique which applies also to mineral and geothermal exploration, to predict seismic hazard or for groundwater monitoring. We aim at developing software package for resistivity inversion, knowing that current numerical methods are not able to manage 3D inversion. We have obtained results based on a Petrov-Galerkin approximation [50], but they are limited to 2D cases. We have thus proposed to reduce the 3D problem by using 1D semi-analytic approximation of Maxwell equations [79]. This work has just started in the framework of a PhD thesis and we hope that it will give us the possibility of imaging 3D problems.

Magique-3D would like to expand its know-how by considering electro-seismic problems which are in the scope of coupling electromagnetic waves with seismic waves. Electro-seismic waves are involved in porous media imaging which is a tricky task because it is based on the coupling of waves with very different wavelengths described by Biot equations and Maxwell equations. Biot equations govern waves in saturated porous media and they represent a complex physical phenomenon involving a slow wave which is very difficult to simulate numerically. In [72], interesting results have been obtained for the simulation of piezoelectric

sensors. They are based on a quasi-static approximation of the Maxwell model coupled with Elastodynamics. Now, we are concerned with the capability of using this model for Geophysical Imaging and we believe that the derivation and/or the analysis of suitable modelings is necessary. Collaborations with Geophysicists are thus mandatory in the prospect of using both experimental and numerical approaches. We would like to collaborate with Prof. C. Bordes and Prof. D. Brito (Laboratory of Complex Fluids and their Reservoirs, CNRS and University of Pau) who have efficient experimental devices for the propagation of electromagnetic waves inside saturated porous media [55]. This collaboration should be easy to organize since Magique-3D has a long-term experience in collaborating with geophysicists. We then believe that we will not need a lot of time to get joint results since we can use our advanced software packages Hou10ni and Montjoie and our colleagues have already obtained data. Electro-seismology is a very challenging research domain for us and we would like to enforce our collaborations with IsTerre (Institute of Earth Science, University of Grenoble) and for that topic with Prof. S. Garambois who is an expert in Electro-seismology [81], [82], [69], [70]. A joint research program could gather Geophysicists from the University of Pau and from IsTerre and Magique-3D. In particular, it would be interesting to compare simulations performed with Hou10ni, Montjoie, with the code developed by Prof. S. Garambois and to use experimental simulations for validation.

3.3. Supercomputing for Helmholtz problems

Probing invisible with harmonic equations is a need for many scientists and it is also a topic offering a wealth of interesting problems for mathematicians. It is well-known that Helmholtz equations discretization is very sensitive to the frequency scale which can be wide-ranging for some applications. For example, depth imaging is searching for deeper layers which may contain hydrocarbons and frequencies must be of a few tens of Hertz with a very low resolution. If it is to detect hidden objects, the depth of the explored region does not exceed a few tens of meters and frequencies close to the kiloHertz are used. High performing numerical methods should thus be stable for a widest as possible frequency range. In particular, these methods should minimize phenomena of numerical pollution that generate errors which increase faster with frequency than with the inverse of space discretization step. As a consequence, there is a need of mesh refinement, in particular at high frequency.

During the period 2010-2014, the team has worked extensively on high order discontinuous Galerkin (DG) methods. Like standard Finite Element Methods, they are elaborated with polynomial basis functions and they are very popular because they are defined locally for each element. It is thus easy to use basis polynomial functions with different degrees and this shows the perfect flexibility of the approximation in case of heterogeneous media including homogeneous parts. Indeed, low degree basis functions can be used in heterogeneous regions where a fine grid is necessary while high degree polynomials can be used for coarse elements covering homogeneous parts. In particular, Magique-3D has developed Hou10ni that solves harmonic wave equations with DG methods and curved elements. We found that both the effects of pollution and dispersion, which are very significant when a conventional finite element method is used, are limited [57]. However, bad conditioning is persisting and reliability of the method is not guaranteed when the coefficients vary considerably. In addition, the number of unknowns of the linear system is too big to hope to solve a realistic 3D problem. So it is important to develop approximation methods that require fewer degrees of freedom. Magique-3D wishes to invest heavily in the development of new approximation methods for harmonic wave equations. It is a difficult subject for which we want to develop different tasks, in collaboration with academic researchers with whom we are already working or have established contacts. Research directions that we would like to follow are the following.

First, we will continue our long-term collaboration with Prof. Rabia Djellouli. We want to continue to work on hybrid finite element methods that rely on basis functions composed of plane waves and polynomials. These methods have demonstrated good resistance to the phenomenon of numerical pollution [51], [52], but their capability of solving industrial problems has not been illustrated. This is certainly due to the absence of guideline for choosing the plane waves. We are thus currently working on the implementation of a methodology that makes the choice of plane waves automatic for a given simulation (fixed propagation domain, data source, etc.). This is up-front investigation and there is certainly a lot of remaining work before

being applied to geophysical imaging. But it gives the team the opportunity to test new ideas while remaining in contact with potential users of the methods.

Then we want to work with Prof. A. Bendali on developing methods of local integral equations which allow calculation of numerical fluxes on the edges of elements. One could then use these fluxes in a DG method for reconstructing the solution throughout the volume of calculation. This research is motivated by recent results which illustrate the difficulties of the existing methods which are not always able to approximate the propagating modes (plane waves) and the evanescent modes (polynomials) that may coexist, especially when one considers realistic applications. Integral equations are direct tools for computing fluxes and they are known for providing very good accuracy. They thus should help to improve the quality of approximation of DG methods which are fully flux-dependent. In addition, local integral equations would limit calculations at the interfaces, which would have the effect of limiting the number of unknowns generally high, especially for DG methods. Again, it is a matter of long-term research which success requires a significant amount of mathematical analysis, and also the development of non-trivial code.

To limit the effects of pollution and dispersion is not the only challenge that the team wants to tackle. Our experience alongside Total has made us aware of the difficulties in constructing meshes that are essential to achieve our simulations. There are several teams at Inria working on mesh generation and we are in contact with them, especially with Gamma3 (Paris-Rocquencourt Research Center). These teams develop meshes increasingly sophisticated to take account of the constraints imposed by realistic industrial benchmarks. But in our opinion, issues which are caused by the construction of meshes are not the only downside. Indeed, we have in mind to solve inverse problems and in this case it is necessary to mesh the domain at each iteration of Newton-type solver. It is therefore interesting to work on methods that either do not use mesh or rely on meshes which are very easy to construct. Regarding meshless methods, we have begun a collaboration with Prof. Djellouli which allowed us to propose a new approach called Mesh-based Frontier Free Formulation (MF3). The principle of this method is the use of fundamental solutions of Helmholtz equations as basic functions. One can then reduce the volumic variational formulation to a surfacic variational formulation which is close to an integral equation, but which does not require the calculation of singularities. The results are very promising and we hope to continue our study in the context of the application to geophysical imaging. An important step to validate this method will be particularly its extension to 3D because the results we have achieved so far are for 2D problems.

Keeping in mind the idea of limiting the difficulties of mesh, we want to study the method of virtual elements. This method attracts us because it relies on meshes that can be made of arbitrarily-shaped polygon and meshes should thus be fairly straightforward. Existing works on the subject have been mainly developed by the University of Pavia, in collaboration with Los Alamos National Laboratory [54], [61], [60], [58], [62]. None of them mentions the feasibility of the method for industrial applications and to our knowledge, there are no results on the method of virtual elements applied to the wave equations. First, we aim at applying the method described in [59] to the scalar Helmholtz equation and explore opportunities to use discontinuous elements within this framework. Then hp-adaptivity could be kept, which is particularly interesting for wave propagation in heterogeneous media.

DG methods are known to require a lot of unknowns that can exceed the limits accepted by the most advanced computers. This is particularly true for harmonic wave equations that require a large number of discretization points, even in the case of a conventional finite element method. We therefore wish to pursue a research activity that we have just started in collaboration with the project-team Nachos (Sophia-Antipolis Méditerranée Research Center). In order to reduce the number of degrees of freedom, we are interested in "hybrid mixed" Discontinuous Galerkin methods that provides a two-step procedure for solving the Helmholtz equations [73], [77], [76]. First, Lagrange multipliers are introduced to represent the flux of the numerical solution through the interface (edge or face) between two elements. The Lagrange multipliers are solution to a linear system which is constructed locally element by element. The number of degrees of freedom is then strongly reduced since for a standard DG method, there is a need of considering unknowns including volumetric values inside the element. And obviously, the gain is even more important when the order of the element is high. Next, the solution is reconstructed from the values of the multipliers and the cost of this step is negligible since it only requires inverting small-sized matrices. We have obtained promising results in the framework of the PhD

thesis of Marie Bonnasse-Gahot and we want to apply it to the simulation of complex phenomena such as the 3D viscoelastic wave propagation.

Obviously, the success of all these works depends on our ability to consider realistic applications such as wave propagation in the Earth. And in these cases, it is quite possible that even if we manage to develop accurate less expensive numerical methods, the solution of inverse problems will still be computationally intensive. It is thus absolutely necessary that we conduct our research by taking advantage of the latest advances in high-performance computing. We have already initiated discussions with the project team HIEPACS (Bordeaux Sud-Ouest research Center) to test the performance of the latest features of Mumps <http://mumps.enseciht.fr/>, such as Low Rank Approximation or adaptation to hybrid CPU / GPU architectures and to Intel Xeon Phi, on realistic test cases. We are also in contact with the team Algorithm at Cerfacs (Toulouse) for the development of local integral equations solvers. These collaborations are essential for us and we believe that they will be decisive for the simulation of three-dimensional elasto-dynamic problems. However, our scientific contribution will be limited in this area because we are not experts in HPC.

3.4. Hybrid time discretizations of high-order

Most of the meshes we consider are composed of cells greatly varying in size. This can be due to the physical characteristics (propagation speed, topography, ...) which may require to refine the mesh locally, very unstructured meshes can also be the result of dysfunction of the mesher. For practical reasons which are essentially guided by the aim of reducing the number of matrix inversions, explicit schemes are generally privileged. However, they work under a stability condition, the so-called Courant Friedrichs Lewy (CFL) condition which forces the time step being proportional to the size of the smallest cell. Then, it is necessary to perform a huge number of iterations in time and in most of the cases because of a very few number of small cells. This implies to apply a very small time step on grids mainly composed of coarse cells and thus, there is a risk of creating numerical dispersion that should not exist. However, this drawback can be avoided by using low degree polynomial basis in space in the small meshes and high degree polynomials in the coarse meshes. By this way, it is possible to relax the CFL condition and in the same time, the dispersion effects are limited. Unfortunately, the cell-size variations are so important that this strategy is not sufficient. One solution could be to apply implicit and unconditionally stable schemes, which would obviously free us from the CFL constraint. Unfortunately, these schemes require inverting a linear system at each iteration and thus needs huge computational burden that can be prohibitive in 3D. Moreover, numerical dispersion may be increased. Then, as second solution is the use of local time stepping strategies for matching the time step to the different sizes of the mesh. There are several attempts [66], [63], [80], [75], [68] and Magique 3D has proposed a new time stepping method which allows us to adapt both the time step and the order of time approximation to the size of the cells. Nevertheless, despite a very good performance assessment in academic configurations, we have observed to our detriment that its implementation inside industrial codes is not obvious and in practice, improvements of the computational costs are disappointing, especially in a HPC framework. Indeed, the local time stepping algorithm may strongly affect the scalability of the code. Moreover, the complexity of the algorithm is increased when dealing with lossy media [71].

Recently, Dolean *et al* [67] have considered a novel approach consisting in applying hybrid schemes combining second order implicit schemes in the thin cells and second order explicit discretization in the coarse mesh. Their numerical results indicate that this method could be a good alternative but the numerical dispersion is still present. It would then be interesting to implement this idea with high-order time schemes to reduce the numerical dispersion. The recent arrival in the team of J. Chabassier should help us to address this problem since she has the expertise in constructing high-order implicit time scheme based on energy preserving Newmark schemes [64]. We propose that our work be organized around the two following tasks. The first one is the extension of these schemes to the case of lossy media because applying existing schemes when there is attenuation is not straightforward. This is a key issue because there is artificial attenuation when absorbing boundary conditions are introduced and if not, there are cases with natural attenuation like in viscoelastic media. The second one is the coupling of high-order implicit schemes with high-order explicit schemes. These two tasks can be first completed independently, but the ultimate goal is obviously to couple the schemes for lossy media. We will consider two strategies for the coupling. The first one will be based on the method

proposed by Dolean *et al*, the second one will consist in using Lagrange multiplier on the interface between the coarse and fine grids and write a novel coupling condition that ensures the high order consistency of the global scheme. Besides these theoretical aspects, we will have to implement the method in industrial codes and our discretization methodology is very suitable for parallel computing since it involves Lagrange multipliers. We propose to organize this task as follows. There is first the crucial issue of a systematic distribution of the cells in the coarse/explicit and in the fine/implicit part. Based on our experience on local time stepping, we claim that it is necessary to define a criterion which discriminates thin cells from coarse ones. Indeed, we intend to develop codes which will be used by practitioners, in particular engineers working in the production department of Total. It implies that the code will be used by people who are not necessarily experts in scientific computing. Considering real-world problems means that the mesh will most probably be composed of a more or less high number of subsets arbitrarily distributed and containing thin or coarse cells. Moreover, in the prospect of solving inverse problems, it is difficult to assess which cells are thin or not in a mesh which varies at each iteration.

Another important issue is the load balancing that we can not avoid with parallel computing. In particular, we will have to choose one of these two alternatives: dedicate one part of processors to the implicit computations and the other one to explicit calculus or distribute the resolution with both schemes on all processors. A collaboration with experts in HPC is then mandatory since we are not expert in parallel computing. We will thus continue to collaborate with the team-projects Hiepacs and Runtime with whom we have a long-term experience of collaborations.

In the future, we aim at enlarging the application range of implicit schemes. The idea will be to use the degrees of freedom offered by the implicit discretization in order to tackle specific difficulties that may appear in some systems. For instance, in systems involving several waves (as P and S waves in porous elastic media, or coupled wave problems as previously mentioned) the implicit parameter could be adapted to each wave and optimized in order to reduce the computational cost. More generally, we aim at reducing numeric bottlenecks by adapting the implicit discretization to specific cases.

SERENA Project-Team

3. Research Program

3.1. Multiphysics coupling

Within our project, we start from the conception and analysis of *models* based on *partial differential equations* (PDEs). Already at the PDE level, we address the question of *coupling* of different models; examples are that of simultaneous fluid flow in a discrete network of two-dimensional *fractures* and in the surrounding three-dimensional porous medium, or that of interaction of a compressible flow with the surrounding elastic *deformable structure*. The key physical characteristics need to be captured, whereas existence, uniqueness, and continuous dependence on the data are minimal analytic requirements that we seek to satisfy. At the modeling stage, we also develop model-order reduction techniques, such as the use of reduced basis techniques or proper generalized decompositions, to tackle evolutive problems, in particular in the nonlinear case.

3.2. Structure-preserving discretizations and discrete element methods

We consequently design *numerical methods* for the devised model. Traditionally, we have worked in the context of finite element, finite volume, mixed finite element, and discontinuous Galerkin methods. Novel classes of schemes enable the use of general *polygonal* and *polyhedral meshes* with *nonmatching interfaces*, and we develop them in response to a high demand from our industrial partners (namely EDF, CEA, and IFP Energies Nouvelles). Our requirement is to derive *structure-preserving* methods, i.e., methods that mimic at the discrete level fundamental properties of the underlying PDEs, such as conservation principles and preservation of invariants. Here, the theoretical questions are closely linked to *differential geometry* for the lowest-order schemes for the Navier–Stokes equations and to the recently-devised theory of gradient schemes for discrete element methods applied to elasto-plasticity. For the schemes we develop, we study existence, uniqueness, and stability questions, and derive a priori convergence estimates. Our special interest is in higher-order methods like the hybrid high-order method, which have recently begun to receive significant attention. Even though their use in practice may not be immediate, we believe that they belong to the future generation of numerical methods for industrial simulations.

3.3. Domain decomposition and Newton–Krylov (multigrid) solvers

We next concentrate an intensive effort on the development and analysis of efficient solvers for the systems of nonlinear algebraic equations that result from the above discretizations. We have in the past developed *Newton–Krylov solvers* like the adaptive inexact Newton method, and we place a particular emphasis on *parallelization* achieved via the *domain decomposition* method. Here we traditionally specialize in *Robin transmission conditions*, where an optimized choice of the parameter has already shown speed-ups in orders of magnitude in terms of the number of domain decomposition iterations in model cases. We concentrate in the SERENA project on adaptation of these algorithms to the above novel discretization schemes, on the optimization of the free Robin parameter for challenging situations, and also on the use of the Ventcell transmission conditions. Another feature is the use of such algorithms in time-dependent problems in *space-time* domain decomposition that we have recently pioneered. This allows the use of different time steps in different parts of the computational domain and turns out to be particularly useful in porous media applications, where the amount of diffusion (permeability) varies abruptly, so that the evolution speed varies significantly from one part of the computational domain to another. Our new theme here are *Newton–multigrid solvers*, where the geometric multigrid solver is *tailored* to the specific problem under consideration and to the specific numerical method, with problem- and discretization-dependent restriction, prolongation, and smoothing. This in particular yields mass balance at each iteration step, a highly demanded feature in most of the target applications. The solver itself is then *adaptively steered* at each execution step by an a posteriori error estimate.

3.4. Reliability by a posteriori error control

The fourth part of our theoretical efforts goes towards guaranteeing the results obtained at the end of the numerical simulation. Here a key ingredient is the development of rigorous *a posteriori estimates* that make it possible to estimate in a fully computable way the error between the unknown exact solution and its numerical approximation. Our estimates also allow to distinguish the different *components* of the overall *error*, namely the errors coming from modeling, from the discretization scheme, from the nonlinear (Newton) solver, and from the linear algebraic (Krylov, domain decomposition, multigrid) solver. A new concept here is that of *local stopping criteria*, where all the error components are balanced locally within each computational mesh element. This naturally connects all parts of the numerical simulation process and gives rise to novel *fully adaptive algorithms*. We shall then address theoretically the question of convergence of the new algorithms and prove their numerical quasi-optimality, meaning that they need, up to a generic constant, the smallest possible number of degrees of freedom to achieve the given accuracy. We in particular seek to prove a guaranteed error reduction in terms of the number of degrees of freedom.

3.5. Safe and correct programming

Finally, we concentrate on the issue of computer implementation of scientific computing programs. Increasing complexity of algorithms for modern scientific computing makes it a major challenge to implement them in the traditional imperative languages popular in the community. As an alternative, the computer science community provides theoretically sound tools for *safe* and *correct programming*. We explore here the use of these tools to design generic solutions for the implementation of the class of scientific computing software that we deal with. Our focus ranges from high-level programming via *functional programming* with OCAML through safe and easy parallelism via *skeleton parallel programming* with SKLML to proofs of correctness of numerical algorithms and programs via *mechanical proofs* with CoQ.

STEPP Project-Team

3. Research Program

3.1. Development of numerical systemic models (economy / society /environment) at local scales

The problem we consider is intrinsically interdisciplinary: it draws on social sciences, ecology or science of the planet. The modeling of the considered phenomena must take into account many factors of different nature which interact with varied functional relationships. These heterogeneous dynamics are *a priori* nonlinear and complex: they may have saturation mechanisms, threshold effects, and may be density dependent. The difficulties are compounded by the strong interconnections of the system (presence of important feedback loops) and multi-scale spatial interactions. Environmental and social phenomena are indeed constrained by the geometry of the area in which they occur. Climate and urbanization are typical examples. These spatial processes involve proximity relationships and neighborhoods, like for example, between two adjacent parcels of land, or between several macroscopic levels of a social organization. The multi-scale issues are due to the simultaneous consideration in the modeling of actors of different types and that operate at specific scales (spatial and temporal). For example, to properly address biodiversity issues, the scale at which we must consider the evolution of rurality is probably very different from the one at which we model the biological phenomena.

In this context, to develop flexible integrated systemic models (upgradable, modular, ...) which are efficient, realistic and easy to use (for developers, modelers and end users) is a challenge in itself. What mathematical representations and what computational tools to use? Nowadays many tools are used: for example, cellular automata (e.g. in the LEAM model), agent models (e.g. URBANSIM⁰), system dynamics (e.g. World3), large systems of ordinary equations (e.g. equilibrium models such as TRANUS), and so on. Each of these tools has strengths and weaknesses. Is it necessary to invent other representations? What is the relevant level of modularity? How to get very modular models while keeping them very coherent and easy to calibrate? Is it preferable to use the same modeling tools for the whole system, or can we freely change the representation for each considered subsystem? How to easily and effectively manage different scales? (difficulty appearing in particular during the calibration process). How to get models which automatically adapt to the granularity of the data and which are always numerically stable? (this has also a direct link with the calibration processes and the propagation of uncertainties). How to develop models that can be calibrated with reasonable efforts, consistent with the (human and material) resources of the agencies and consulting firms that use them?

Before describing our research axes, we provide a brief overview of the types of models that we are or will be working with. As for LUTI (Land Use and Transportation Integrated) modeling, we have been using the TRANUS model since the start of our group. It is the most widely used LUTI model, has been developed since 1982 by the company Modelistica, and is distributed *via* Open Source software. TRANUS proceeds by solving a system of deterministic nonlinear equations and inequalities containing a number of economic parameters (e.g. demand elasticity parameters, location dispersion parameters, etc.). The solution of such a system represents an economic equilibrium between supply and demand.

On the other hand, the scientific domains related to ecosystem services and ecological accounting are much less mature than the one of urban economy from a modelling point of view (as a consequence of our more limited knowledge of the relevant complex processes and/or more limited available data). Nowadays, the community working on ecological accounting develops statistical models based on the enforcement of the mass conservation constraint for accounting for material fluxes through a territorial unit or a supply chain, relying on more or less simple data correlations when the relevant data is missing; the overall modelling makes heavy use of more or less sophisticated linear algebra and constrained optimization techniques. The

⁰<http://www.urbansim.org>

ecosystem service community has been using static models too, but is also developing more sophisticated models based for example on system dynamics, multi-agent type simulations or cellular models. In the ESNET project, STEEP has worked in particular on a land use/ land cover change (LUCC) modelling environments (Dinamica ⁰) which belongs to the category of spatially explicit statistical models.

In the following, our two main research axes are described, from the point of view of applied mathematical development. The domains of application of this research effort is described in the application section, where some details about the context of each field is given.

3.2. Model calibration and validation

The overall calibration of the parameters that drive the equations implemented in the above models is a vital step. Theoretically, as the implemented equations describe e.g. socio-economic phenomena, some of these parameters should in principle be accurately estimated from past data using econometrics and statistical methods like regressions or maximum likelihood estimates, e.g. for the parameters of logit models describing the residential choices of households. However, this theoretical consideration is often not efficient in practice for at least two main reasons. First, the above models consist of several interacting modules. Currently, these modules are typically calibrated independently; this is clearly sub-optimal as results will differ from those obtained after a global calibration of the interaction system, which is the actual final objective of a calibration procedure. Second, the lack of data is an inherent problem.

As a consequence, models are usually calibrated by hand. The calibration can typically take up to 6 months for a medium size LUTI model (about 100 geographic zones, about 10 sectors including economic sectors, population and employment categories). This clearly emphasizes the need to further investigate and at least semi-automate the calibration process. Yet, in all domains STEEP considers, very few studies have addressed this central issue, not to mention calibration under uncertainty which has largely been ignored (with the exception of a few uncertainty propagation analyses reported in the literature).

Besides uncertainty analysis, another main aspect of calibration is numerical optimization. The general state-of-the-art on optimization procedures is extremely large and mature, covering many different types of optimization problems, in terms of size (number of parameters and data) and type of cost function(s) and constraints. Depending on the characteristics of the considered models in terms of dimension, data availability and quality, deterministic or stochastic methods will be implemented. For the former, due to the presence of non-differentiability, it is likely, depending on their severity, that derivative free control methods will have to be preferred. For the latter, particle-based filtering techniques and/or metamodel-based optimization techniques (also called response surfaces or surrogate models) are good candidates.

These methods will be validated, by performing a series of tests to verify that the optimization algorithms are efficient in the sense that 1) they converge after an acceptable computing time, 2) they are robust and 3) that the algorithms do what they are actually meant to. For the latter, the procedure for this algorithmic validation phase will be to measure the quality of the results obtained after the calibration, i.e. we have to analyze if the calibrated model fits sufficiently well the data according to predetermined criteria.

To summarize, the overall goal of this research axis is to address two major issues related to calibration and validation of models: (a) defining a calibration methodology and developing relevant and efficient algorithms to facilitate the parameter estimation of considered models; (b) defining a validation methodology and developing the related algorithms (this is complemented by sensitivity analysis, see the following section). In both cases, analyzing the uncertainty that may arise either from the data or the underlying equations, and quantifying how these uncertainties propagate in the model, are of major importance. We will work on all those issues for the models of all the applied domains covered by STEEP.

3.3. Sensitivity analysis

⁰<http://www.csr.ufmg.br/dinamica/>

A sensitivity analysis (SA) consists, in a nutshell, in studying how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model inputs. It is complementary to an uncertainty analysis, which focuses on quantifying uncertainty in model output. SA's can be useful for several purposes, such as guiding model development and identifying the most influential model parameters and critical data items. Identifying influential model parameters may help in devising metamodels (or, surrogate models) that approximate an original model and may be simulated, calibrated, or analyzed more efficiently. As for detecting critical data items, this may indicate for which type of data more effort must be spent in the data collection process in order to eventually improve the model's reliability. Finally, SA can be used as one means for validating models, together with validation based on historical data (or, put simply, using training and test data) and validation of model parameters and outputs by experts in the respective application area.

The first two applications of SA are linked to model calibration, discussed in the previous section. Indeed, prior to the development of the calibration tools, one important step is to select the significant or sensitive parameters and to evaluate the robustness of the calibration results with respect to data noise (stability studies). This may be performed through a global sensitivity analysis, e.g. by computation of Sobol's indices. Many problems had been to be circumvented e.g. difficulties arising from dependencies of input variables, variables that obey a spatial organization, or switch inputs. We take up on current work in the statistics community on SA for these difficult cases.

As for the third application of SA, model validation, a preliminary task bears on the propagation of uncertainties. Identifying the sources of uncertainties and their nature is crucial to propagate them via Monte Carlo techniques. To make a Monte Carlo approach computationally feasible, it is necessary to develop specific metamodels. Both the identification of the uncertainties and their propagation require a detailed knowledge of the data collection process; these are mandatory steps before a validation procedure based on SA can be implemented. First, we focus on validating LUTI models, starting with the CITiES ANR project: here, an SA consists in defining various land use policies and transportation scenarios and in using these scenarios to test the integrated land use and transportation model. Current approaches for validation by SA consider several scenarios and propose various indicators to measure the simulated changes. We work towards using sensitivity indices based on functional analysis of variance, which allow us to compare the influence of various inputs on the indicators. For example it allow the comparison of the influences of transportation and land use policies on several indicators.

TONUS Team

3. Research Program

3.1. Kinetic models for plasmas

The fundamental model for plasma physics is the coupled Vlasov-Maxwell kinetic model: the Vlasov equation describes the distribution function of particles (ions and electrons), while the Maxwell equations describe the electromagnetic field. In some applications, it may be necessary to take relativistic particles into account, which leads to consider the relativistic Vlasov equation, even if in general, tokamak plasmas are supposed to be non-relativistic. The distribution function of particles depends on seven variables (three for space, three for the velocity and one for time), which yields a huge amount of computations.

To these equations we must add several types of source terms and boundary conditions for representing the walls of the tokamak, the applied electromagnetic field that confines the plasma, fuel injection, collision effects, etc.

Tokamak plasmas possess particular features, which require developing specialized theoretical and numerical tools.

Because the magnetic field is strong, the particle trajectories have a very fast rotation around the magnetic field lines. A full resolution would require a prohibitive amount of calculations. It is then necessary to develop reduced models for large magnetic fields in order to obtain tractable calculations. The resulting model is called a gyrokinetic model. It allows us to reduce the dimensionality of the problem. Such models are implemented in GYSELA and Selalib.

On the boundary of the plasma, the collisions can no more be neglected. Fluid models, such as the MagnetoHydroDynamics (MHD) become again relevant. For the good operation of the tokamak, it is necessary to control MHD instabilities that arise at the plasma boundary. Computing these instabilities requires special implicit numerical discretizations with excellent long time behavior.

In addition to theoretical modelling tools, it is necessary to develop numerical schemes adapted to kinetic, gyrokinetic and fluid models. Three kinds of methods are studied in TONUS: Particle-In-Cell (PIC) methods, semi-Lagrangian and fully Eulerian approaches.

3.1.1. Gyrokinetic models: theory and approximation

In most phenomena where oscillations are present, we can establish a three-model hierarchy: *(i)* the model parameterized by the oscillation period, *(ii)* the limit model and *(iii)* the two-scale model, possibly with its corrector. In a context where one wishes to simulate such a phenomenon where the oscillation period is small and the oscillation amplitude is not small, it is important to have numerical methods based on an approximation of the Two-Scale model. If the oscillation period varies significantly over the domain of simulation, it is important to have numerical methods that approximate properly and effectively the model parameterized by the oscillation period and the Two-Scale model. Implementing Two-Scale Numerical Methods (for instance by Frénod et al. [23]) is based on the numerical approximation of the Two-Scale model. These are called of order 0. A Two-Scale Numerical Method is called of order 1 if it incorporates information from the corrector and from the equation of which this corrector is a solution. If the oscillation period varies between very small values and values of order 1, it is necessary to have new types of numerical schemes (Two-Scale Asymptotic Preserving Schemes of order 1 or TSAPS) that preserve the asymptotics between the model parameterized by the oscillation period and the Two-Scale model with its corrector. A first work in this direction has been initiated by Crouseilles et al. [22].

3.1.2. Semi-Lagrangian schemes

The Strasbourg team has a long and recognized experience in numerical methods of Vlasov-type equations. We are specialized in both particle and phase space solvers for the Vlasov equation: Particle-in-Cell (PIC) methods and semi-Lagrangian methods. We also have a long-standing collaboration with the CEA of Cadarache for the development of the GYSELA software for gyrokinetic tokamak plasmas.

The Vlasov and the gyrokinetic models are partial differential equations that express the transport of the distribution function in the phase space. In the original Vlasov case, the phase space is the six-dimension position-velocity space. For the gyrokinetic model, the phase space is five-dimensional because we consider only the parallel velocity in the direction of the magnetic field and the gyrokinetic angular velocity instead of three velocity components.

A few years ago, Eric Sonnendrücker and his collaborators introduced a new family of methods for solving transport equations in the phase space. This family of methods are the semi-Lagrangian methods. The principle of these methods is to solve the equation on a grid of the phase space. The grid points are transported with the flow of the transport equation for a time step and interpolated back periodically onto the initial grid. The method is then a mix of particle Lagrangian methods and Eulerian methods. The characteristics can be solved forward or backward in time leading to the Forward Semi-Lagrangian (FSL) or Backward Semi-Lagrangian (BSL) schemes. Conservative schemes based on this idea can be developed and are called Conservative Semi-Lagrangian (CSL).

GYSELA is a 5D full gyrokinetic code based on a classical backward semi-Lagrangian scheme (BSL) [27] for the simulation of core turbulence that has been developed at CEA Cadarache in collaboration with our team [24].

More recently, we have started to apply the Semi-Lagrangian methods to more general kinetic equations. Indeed, most of the conservation laws of physics can be represented by a kinetic model with a small set of velocities and relaxation source terms [7]. Compressible fluids or MHD equations have such representations. Semi-Lagrangian methods then become a very appealing and efficient approach for solving these equations.

3.1.3. PIC methods

Historically PIC methods have been very popular for solving the Vlasov equations. They allow solving the equations in the phase space at a relatively low cost. The main disadvantage of this approach is that, due to its random aspect, it produces an important numerical noise that has to be controlled in some way, for instance by regularizations of the particles, or by divergence correction techniques in the Maxwell solver. We have a long-standing experience in PIC methods and we started implementing them in Selalib. An important aspect is to adapt the method to new multicore computers. See the work by Crestetto and Helluy [21].

3.2. Fluid and Reduced kinetic models for plasmas

As already said, kinetic plasmas computer simulations are very intensive, because of the gyrokinetic turbulence. In some situations, it is possible to make assumptions on the shape of the distribution function that simplify the model. We obtain in this way a family of fluid or reduced models.

Assuming that the distribution function has a Maxwellian shape, for instance, we obtain the MagnetoHydro-Dynamic (MHD) model. It is physically valid only in some parts of the tokamak (at the edges for instance). The fluid model is generally obtained from the hypothesis that the collisions between particles are strong.

But the reduction is not necessarily a consequence of collisional effects. Indeed, even without collisions, the plasma may still relax to an equilibrium state over sufficiently long time scales (Landau damping effect).

In the fluid or reduced-kinetic regions, the approximation of the distribution function could require fewer data while still achieving a good representation, even in the collisionless regime.

Therefore, a fluid or a reduced model is a model where the explicit dependency on the velocity variable is removed. In a more mathematical way, we consider that in some regions of the plasma, it is possible to exhibit a (preferably small) set of parameters α that allows us to describe the main properties of the plasma with a generalized “Maxwellian” M . Then

$$f(x, v, t) = M(\alpha(x, t), v).$$

In this case it is sufficient to solve for $\alpha(x, t)$. Generally, the vector α is the solution of a first order hyperbolic system.

Another way to reduce the model is to try to find an abstract kinetic representation with an as small as possible set of kinetic velocities. The kinetic approach has then only a mathematical meaning. It allows solving very efficiently many equations of physics [13].

3.2.1. Numerical schemes

As previously indicated, an efficient method for solving the reduced models is the Discontinuous Galerkin (DG) approach. It is possible to make it of arbitrary order. It requires limiters when it is applied to nonlinear PDEs occurring for instance in fluid mechanics. But the reduced models that we intent to write are essentially linear. The nonlinearity is concentrated in a few coupling source terms.

In addition, this method, when written in a special set of variables, called the entropy variables, has nice properties concerning the entropy dissipation of the model. It opens the door to constructing numerical schemes with good conservation properties and no entropy dissipation, as already used for other systems of PDEs [28], [20], [26], [25].

3.2.2. Matrix-free Implicit schemes

In tokamaks, the reduced model generally involves many time scales. Among these time scales, many of them, associated to the fastest waves, are not relevant. In order to filter them out, it is necessary to adopt implicit solvers in time. When the reduced model is based on a kinetic interpretation, it is possible to construct implicit schemes that do not impose solving costly linear systems. In addition the resulting solver is stable even at very high CFL number [13].

3.3. Electromagnetic solvers

Precise resolution of the electromagnetic fields is essential for proper plasma simulation. Thus it is important to use efficient solvers for the Maxwell systems and its asymptotics: Poisson equation and magnetostatics.

The proper coupling of the electromagnetic solver with the Vlasov solver is also crucial for ensuring conservation properties and stability of the simulation.

Finally, plasma physics implies very different time scales. It is thus very important to develop implicit Maxwell solvers and Asymptotic Preserving (AP) schemes in order to obtain good behavior on long time scales.

3.3.1. Coupling

The coupling of the Maxwell equations to the Vlasov solver requires some precautions. The most important one is to control the charge conservation errors, which are related to the divergence conditions on the electric and magnetic fields. We will generally use divergence correction tools for hyperbolic systems presented for instance in [18] (and the references therein).

3.3.2. Implicit solvers

As already pointed out, in a tokamak, the plasma presents several different space and time scales. It is not possible in practice to solve the initial Vlasov-Maxwell model. It is first necessary to establish asymptotic models by letting some parameters (such as the Larmor frequency or the speed of light) tend to infinity. This is the case for the electromagnetic solver and this requires implementing implicit time solvers in order to efficiently capture the stationary state, the solution of the magnetic induction equation or the Poisson equation.

BIOCORE Project-Team

3. Research Program

3.1. Mathematical and computational methods

BIOCORE's action is centered on the mathematical modeling of biological systems, more particularly of artificial ecosystems, that have been built or strongly shaped by human. Indeed, the complexity of such systems where life plays a central role often makes them impossible to understand, control, or optimize without such a formalization. Our theoretical framework of choice for that purpose is Control Theory, whose central concept is "the system", described by state variables, with inputs (action on the system), and outputs (the available measurements on the system). In modeling the ecosystems that we consider, mainly through ordinary differential equations, the state variables are often population, substrate and/or food densities, whose evolution is influenced by the voluntary or involuntary actions of man (inputs and disturbances). The outputs will be some product that one can collect from this ecosystem (harvest, capture, production of a biochemical product, etc), or some measurements (number of individuals, concentrations, etc). Developing a model in biology is however not straightforward: the absence of rigorous laws as in physics, the presence of numerous populations and inputs in the ecosystems, most of them being irrelevant to the problem at hand, the uncertainties and noise in experiments or even in the biological interactions require the development of dedicated techniques to identify and validate the structure of models from data obtained by or with experimentalists.

Building a model is rarely an objective in itself. Once we have checked that it satisfies some biological constraints (eg. densities stay positive) and fitted its parameters to data (requiring tailor-made methods), we perform a mathematical analysis to check that its behavior is consistent with observations. Again, specific methods for this analysis need to be developed that take advantage of the structure of the model (eg. the interactions are monotone) and that take into account the strong uncertainty that is linked to life, so that qualitative, rather than quantitative, analysis is often the way to go.

In order to act on the system, which often is the purpose of our modeling approach, we then make use of two strong points of Control Theory: 1) the development of observers, that estimate the full internal state of the system from the measurements that we have, and 2) the design of a control law, that imposes to the system the behavior that we want to achieve, such as the regulation at a set point or optimization of its functioning. However, due to the peculiar structure and large uncertainties of our models, we need to develop specific methods. Since actual sensors can be quite costly or simply do not exist, a large part of the internal state often needs to be re-constructed from the measurements and one of the methods we developed consists in integrating the large uncertainties by assuming that some parameters or inputs belong to given intervals. We then developed robust observers that asymptotically estimate intervals for the state variables [61]. Using the directly measured variables and those that have been obtained through such, or other, observers, we then develop control methods that take advantage of the system structure (linked to competition or predation relationships between species in bioreactors or in the trophic networks created or modified by biological control).

3.2. A methodological approach to biology: from genes to ecosystems

One of the objectives of BIOCORE is to develop a methodology that leads to the integration of the different biological levels in our modeling approach: from the biochemical reactions to ecosystems. The regulatory pathways at the cellular level are at the basis of the behavior of the individual organism but, conversely, the external stresses perceived by the individual or population will also influence the intracellular pathways. In a modern "systems biology" view, the dynamics of the whole biosystem/ecosystem emerge from the interconnections among its components, cellular pathways/individual organisms/population. The different scales of size and time that exist at each level will also play an important role in the behavior of the biosystem/ecosystem. We intend to develop methods to understand the mechanisms at play at each level,

from cellular pathways to individual organisms and populations; we assess and model the interconnections and influence between two scale levels (eg., metabolic and genetic; individual organism and population); we explore the possible regulatory and control pathways between two levels; we aim at reducing the size of these large models, in order to isolate subsystems of the main players involved in specific dynamical behaviors.

We develop a theoretical approach of biology by simultaneously considering different levels of description and by linking them, either bottom up (scale transfer) or top down (model reduction). These approaches are used on modeling and analysis of the dynamics of populations of organisms; modeling and analysis of small artificial biological systems using methods of systems biology; control and design of artificial and synthetic biological systems, especially through the coupling of systems.

The goal of this multi-level approach is to be able to design or control the cell or individuals in order to optimize some production or behavior at higher level: for example, control the growth of microalgae via their genetic or metabolic networks, in order to optimize the production of lipids for bioenergy at the photobioreactor level.

CARMEN Project-Team

3. Research Program

3.1. Complex models for the propagation of cardiac action potentials

The contraction of the heart is coordinated by a complex electrical activation process which relies on about a million ion channels, pumps, and exchangers of various kinds in the membrane of each cardiac cell. Their interaction results in a periodic change in transmembrane potential called an action potential. Action potentials in the cardiac muscle propagate rapidly from cell to cell, synchronizing the contraction of the entire muscle to achieve an efficient pump function. The spatio-temporal pattern of this propagation is related both to the function of the cellular membrane and to the structural organization of the cells into tissues. Cardiac arrhythmias originate from malfunctions in this process. The field of cardiac electrophysiology studies the multiscale organization of the cardiac activation process from the subcellular scale up to the scale of the body. It relates the molecular processes in the cell membranes to the propagation process and to measurable signals in the heart and to the electrocardiogram, an electrical signal on the torso surface.

Several improvements of current models of the propagation of the action potential are being developed in the Carmen team, based on previous work [49] and on the data available at IHU LIRYC:

- Enrichment of the current monodomain and bidomain models [49], [59] by accounting for structural heterogeneities of the tissue at an intermediate scale. Here we focus on multiscale analysis techniques applied to the various high-resolution structural data available at the LIRYC.
- Coupling of the tissues from the different cardiac compartments and conduction systems. Here, we develop models that couple 1D, 2D and 3D phenomena described by reaction-diffusion PDEs.

These models are essential to improve our in-depth understanding of cardiac electrical dysfunction. To this aim, we use high-performance computing techniques in order to numerically explore the complexity of these models.

We use these model codes for applied studies in two important areas of cardiac electrophysiology: atrial fibrillation [51] and sudden-cardiac-death (SCD) syndromes [7], [6] [54]. This work is performed in collaboration with several physiologists and clinicians both at IHU Liryc and abroad.

3.2. Simplified models and inverse problems

The medical and clinical exploration of the cardiac electric signals is based on accurate reconstruction of the patterns of propagation of the action potential. The correct detection of these complex patterns by non-invasive electrical imaging techniques has to be developed. This problem involves solving inverse problems that cannot be addressed with the more complex models. We want both to develop simple and fast models of the propagation of cardiac action potentials and improve the solutions to the inverse problems found in cardiac electrical imaging techniques.

The cardiac inverse problem consists in finding the cardiac activation maps or, more generally, the whole cardiac electrical activity, from high-density body surface electrocardiograms. It is a new and a powerful diagnosis technique, which success would be considered as a breakthrough. Although widely studied recently, it remains a challenge for the scientific community. In many cases the quality of reconstructed electrical potential is not adequate. The methods used consist in solving the Laplace equation on the volume delimited by the body surface and the epicardial surface. Our aim is to

- study in depth the dependance of this inverse problem on inhomogeneities in the torso, conductivity values, the geometry, electrode positions, etc., and
- improve the solution to the inverse problem by using new regularization strategies, factorization of boundary value problems, and the theory of optimal control, both in the quasistatic and in the dynamic contexts.

Of course we will use our models as a basis to regularize these inverse problems. We will consider the following strategies:

- using complete propagation models in the inverse problem, like the bidomain equations, for instance in order to localize electrical sources;
- constructing families of reduced-order models using e.g. statistical learning techniques, which would accurately represent some families of well-identified pathologies; and
- constructing simple models of the propagation of the activation front, based on eikonal or level-set equations, but which would incorporate the representation of complex activation patterns.

Additionally, we will need to develop numerical techniques dedicated to our simplified eikonal/level-set equations.

3.3. Numerical techniques

We want the numerical simulations of the previous direct or inverse models to be efficient and reliable with respect to the needs of the medical community. They should qualify and guarantee the accuracy and robustness of the numerical techniques and the efficiency of the resolution algorithms.

Based on previous work on solving the monodomain and bidomain equations [4], [5], [8], [1], we will focus on

- High-order numerical techniques with respect to the variables with physiological meaning, like velocity, AP duration and restitution properties.
- Efficient, dedicated preconditioning techniques coupled with parallel computing.

Existing simulation tools used in our team rely, among others, on mixtures of explicit and implicit integration methods for ODEs, hybrid MPI-OpenMP parallelization, algebraic multigrid preconditioning, and a BiCGStab algorithm with adaptations to retain numerical accuracy while handling large underdetermined systems.

3.4. Cardiac Electrophysiology at the Microscopic Scale

Numerical models of whole-heart physiology are based on the approximation of a perfect muscle using homogenisation methods. However, due to aging and cardiomyopathies, the cellular structure of the tissue changes. These modifications can give rise to life-threatening arrhythmias. For our research on this subject and with cardiologists of the IHU LIRYC Bordeaux, we aim to design and implement models that describe the strong heterogeneity of the tissue at the cellular level and to numerically explore the mechanisms of these diseases.

The literature on this type of model is still very limited [63]. Existing models are two-dimensional [55] or limited to idealized geometries, and use a linear (purely resistive) behaviour of the gap-junction channels that connect the cells. We propose a three-dimensional approach using realistic cellular geometry (figure 1), nonlinear gap-junction behaviour, and a numerical approach that can scale to hundreds of cells while maintaining a sub-micrometer spatial resolution (10 to 100 times smaller than the size of a cardiomyocyte) [28] [48], [47].

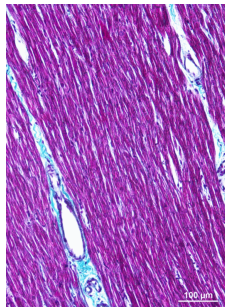
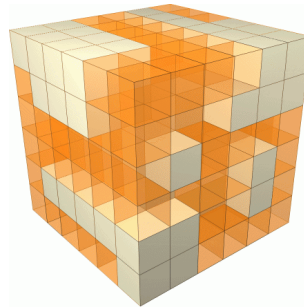
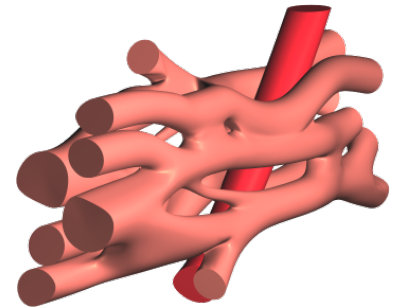
**A****B****C**

Figure 1. **A:** The cardiac muscle consists of a branching network of elongated muscle cells, interspersed with other structures. Sheets of connective tissue (blue) can grow between the muscle cells and become pathogenic. **B:** Current models can only represent such alterations in a coarse way by replacing model elements with different types; each cube in this illustration would represent hundreds of cells. **C:** This hand-crafted example illustrates the type of geometric model we are experimenting with. Each cell is here represented by hundreds of elements.

DRACULA Project-Team

3. Research Program

3.1. Mixed-effect models and statistical approaches

Most of biological and medical data our team has to deal with consist in time series of experimental measurements (cell counts, gene expression level, etc.). The intrinsic variability of any biological system complicates its confrontation to models. The trivial use of means, eliminating the data variance, is but a second-best solution. Furthermore, the amount of data that can be experimentally generated often limits the use of classical mathematical approaches because model's identifiability or parameter identifiability cannot be obtained. In order to overcome this issue and to efficiently take advantage of existing and available data, we plan to use mixed effect models for various applications (for instance: leukemia treatment modeling, immune response modeling). Such models were initially developed to account for individual behaviors within a population by characterizing distributions of parameter values instead of a unique parameter value. We plan to use those approaches both within that frame (for example, taking into account longitudinal studies on different patients, or different mice) but also to extend its validity in a different context: we will consider different ex vivo experiments as being "different individuals": this will allow us to make the most of the experience-to-experience variations.

Such approaches need expertise in statistics to be correctly implemented, and we will rely on the presence of Céline Vial in the team to do so. Céline Vial is an expert in applied statistics and her experience already motivated the use of better statistical methods in various research themes. The increasing use of single cell technologies in biology make such approaches necessary and it is going to be critical for the project to acquire such skills.

3.2. Development of a simulation platform

We have put some effort in developing the *SiMuScale* platform, a software coded in C++ dedicated to exploring multiscale population models, since 2014. In order to answer the challenges of multi-scale modeling it is necessary to possess an all-purpose, fast and flexible modeling tool, and *SiMuScale* is the choice we made. Since it is based on a core containing the simulator, and on plug-ins that contain the biological specifications of each cell, this software will make it easier for members of the team – and potentially other modelers – to focus on the model and to capitalize on existing models, which all share the same framework and are compatible with each other. Within the next four years, *SiMuScale* should be widely accessible and daily used in the team for multi-scale modeling. It will be developed into a real-case context, the modeling of the hematopoietic stem cell niche, in collaboration with clinicians (Eric Solary, INSERM) and physicists (Bertrand Laforge, UPMC).

3.3. Mathematical and computational modeling

Multi-scale modeling of hematopoiesis is one of the key points of the project that has started in the early stage of the Dracula team. Investigated by the team members, it took many years of close discussion with biologists to get the best understanding of the key role played by the most important molecules, hormones, kinase cascade, cell communication up to the latest knowledge. An approach that we used is based on hybrid discrete-continuous models, where cells are considered as individual objects, intracellular regulatory networks are described with ordinary differential equations, extracellular concentrations with diffusion or diffusion-convection equations (see Figure 1). These modeling tools require the expertise of all team members to get the most qualitative satisfactory model. The obtained models will be applied particularly to describe normal and pathological hematopoiesis as well as immune response.

3.4. From hybrid dynamics to continuum mechanics

Hybrid discrete-continuous methods are well adapted to describe biological cells. However, they are not appropriate for the qualitative investigation of the corresponding phenomena. Therefore, hybrid model approach should be combined with continuous models. If we consider cell populations as a continuous medium, then cell concentrations can be described by reaction-diffusion systems of equations with convective terms. The diffusion terms correspond to a random cell motion and the reaction terms to cell proliferation, differentiation and death. We will continue our studies of stability, nonlinear dynamics and pattern formation. Theoretical investigations of reaction-diffusion models will be accompanied by numerical simulations and will be applied to study cell population dynamic.

3.5. Structured partial differential equations

Hyperbolic problems are also of importance when describing cell population dynamics. They are structured transport partial differential equations, in which the structure is a characteristic of the considered population, for instance age, size, maturity, etc. In the scope of multi-scale modeling, protein concentrations as structure variables can precisely indicate the nature of cellular events cells undergo (differentiation, apoptosis), by allowing a representation of cell populations in a multi-dimensional space. Several questions are still open in the study of this problem, yet we will continue our analysis of these equations by focusing in particular on the asymptotic behavior of the system (stability, oscillations) and numerical simulations.

3.6. Delay differential equations

The use of age structure in PDE often leads to a reduction (by integration over the age variable) to delay differential equations. Delay differential equations are particularly useful for situations where the processes are controlled through feedback loops acting after a certain time. For example, in the evolution of cell populations the transmission of control signals can be related to some processes as division, differentiation, maturation, apoptosis, etc. Delay differential equations offer good tools to study the behavior of the systems. Our main investigation will be the effect of perturbations of the parameters, as cell cycle duration, apoptosis, differentiation, self-renewal, etc., on the behavior of the system, in relation for instance with some pathological situations. The mathematical analysis of delay differential equations is often complicated and needs the development of new criteria to be performed.

3.7. Multi-scale modeling of the immune response

The main objective of this part is to develop models that make it possible to investigate the dynamics of the adaptive CD8 T cell immune response, and in particular to focus on the consequences of early molecular events on the cellular dynamics few days or weeks later: this would help developing predictive tools of the immune response in order to facilitate vaccine development and reduce costs. This work requires a close and intensive collaboration with immunologist partners.

We recently published a model of the CD8 T cell immune response characterizing differentiation stages, identified by biomarkers, able to predict the quantity of memory cells from early measurements ([17]). In parallel, we improved our multiscale model of the CD8 T cell immune response, by implementing a full differentiation scheme, from naïve to memory cells, based on a limited set of genes and transcription factors.

Our first task will be to infer an appropriate gene regulatory network (GRN) using single cell data analysis (generate transcriptomics data of the CD8 T cell response to diverse pathogens), the previous biomarkers we identified and associated to differentiation stages, as well as piecewise-deterministic Markov processes (Ulysse Herbach's PhD thesis, ongoing).

Our second task will be to update our multiscale model by first implementing the new differentiation scheme we identified ([17]), and second by embedding CD8 T cells with the GRN obtained in our first task (see above). This will lead to a multi-scale model incorporating description of the CD8 T cell immune response both at the molecular and the cellular levels (Simon Girel's PhD thesis, ongoing).

In order to further develop our multiscale model, we will consider an agent-based approach for the description of the cellular dynamics. Yet, such models, coupled to continuous models describing GRN dynamics, are computationally expensive, so we will focus on alternative strategies, in particular on descriptions of the cellular dynamics through both continuous and discrete models, efficiently coupled. Using discrete models for low cell numbers and continuous (partial differential equations) models for large cell numbers, with appropriate coupling strategies, can lead to faster numerical simulations, and consequently can allow performing intense parameter estimation procedures that are necessary to validate models by confronting them to experimental data, both at the molecular and cellular scales.

The final objective will be to capture CD8 T cell responses in different immunization contexts (different pathogens, tumor) and to predict cellular outcomes from molecular events.

3.8. Dynamical network inference from single-cell data

Up to now, all of our multiscale models have incorporated a dynamical molecular network that was build “by hand” after a thorough review of the literature. It would be highly valuable to infer it directly from gene expression data. However, this remains very challenging from a methodological point of view. We started exploring an original solution for such inference by using the information contained within gene expression distributions. Such distributions can be acquired through novel techniques where gene expression levels are quantified at the single cell level. We propose to view the inference problem as a fitting procedure for a mechanistic gene network model that is inherently stochastic and takes not only protein, but also mRNA levels into account. This approach led to very encouraging results [21] and we will actively pursue in that direction, especially in the light of the foreseeable explosion of single cell data.

3.9. Leukemia modeling

Imatinib and other tyrosine kinase inhibitors (TKIs) have marked a revolution in the treatment of Chronic Myelogenous Leukemia (CML). Yet, most patients are not cured, and must take their treatment for life. Deeper mechanistic understanding could improve TKI combination therapies to better control the residual leukemic cell population. In a collaboration with the Hospital Lyon Sud and the University of Maryland, we have developed mathematical models that integrate CML and an autologous immune response ([11], [12] and [34]). These studies have lent theoretical support to the idea that the immune system plays a rôle in maintaining remission over long periods. Our mathematical model predicts that upon treatment discontinuation, the immune system can control the disease and prevent a relapse. There is however a possibility for relapse via a sneak-though mechanism [11]. Research in the next four years will focus in the Phase III PETALS trial. In the PETALS trial (<https://clinicaltrials.gov/ct2/show/NCT02201459>), the second generation TKI Nilotinib is combined with Peg-IFN, an interferon that is thought to enhance the immune response. We plan to: 1) Adapt the model to take into account the early dynamics (first three months). 2) Use a mixed-effect approach to analyse the effect of the combination, and find population and individual parameters related to treatment efficacy and immune system response. 3) Optimise long-term treatment strategies to reduce or cease treatment and make personalised predictions based on mixed-effect parameters, to minimise the long-term probability of relapse.

M3DISIM Project-Team

3. Research Program

3.1. Multi-scale modeling and coupling mechanisms for biomechanical systems, with mathematical and numerical analysis

Over the past decade, we have laid out the foundations of a multi-scale 3D model of the cardiac mechanical contraction responding to electrical activation. Several collaborations have been crucial in this enterprise, see below references. By integrating this formulation with adapted numerical methods, we are now able to represent the whole organ behavior in interaction with the blood during complete heart beats. This subject was our first achievement to combine a deep understanding of the underlying physics and physiology and our constant concern of proposing well-posed mathematical formulations and adequate numerical discretizations. In fact, we have shown that our model satisfies the essential thermo-mechanical laws, and in particular the energy balance, and proposed compatible numerical schemes that – in consequence – can be rigorously analyzed, see [4]. In the same spirit, we have recently formulated a poromechanical model adapted to the blood perfusion in the heart, hence precisely taking into account the large deformation of the mechanical medium, the fluid inertia and moving domain, and so that the energy balance between fluid and solid is fulfilled from the model construction to its discretization, see [5].

3.2. Inverse problems with actual data – Fundamental formulation, mathematical analysis and applications

A major challenge in the context of biomechanical modeling – and more generally in modeling for life sciences – lies in using the large amount of data available on the system to circumvent the lack of absolute modeling ground truth, since every system considered is in fact patient-specific, with possibly non-standard conditions associated with a disease. We have already developed original strategies for solving this particular type of inverse problems by adopting the observer stand-point. The idea we proposed consists in incorporating to the classical discretization of the mechanical system an estimator filter that can use the data to improve the quality of the global approximation, and concurrently identify some uncertain parameters possibly related to a diseased state of the patient, see [7], [8], [9]. Therefore, our strategy leads to a coupled model-data system solved similarly to a usual PDE-based model, with a computational cost directly comparable to classical Galerkin approximations. We have already worked on the formulation, the mathematical and numerical analysis of the resulting system – see [3] – and the demonstration of the capabilities of this approach in the context of identification of constitutive parameters for a heart model with real data, including medical imaging, see [1].

MAMBA Project-Team

3. Research Program

3.1. Introduction

Data and image analysis, statistical, ODEs, PDEs, and agent-based approaches are used either individually or in combination, with a strong focus on PDE analysis and agent-based approaches. Mamba was created in January 2014, as a continuation of the BANG project-team, that had been headed by Benoît Perthame from 2003-2013, and in the last years increasingly broaden its subjects as its individuals develop their own research agendas. It aims at developing models, simulations and numerical algorithms to solve questions from life sciences involving dynamics of phenomena encountered in biological systems such as protein intra-cellular spatio-temporal dynamics, cell motion, early embryonic development, multicellular growth, wound healing and liver regeneration, cancer evolution, healthy and tumour growth control by pharmaceuticals, protein polymerisation occurring in neurodegenerative disorders, etc.

Another guideline of our project is to remain close to the most recent questions of experimental biology or medicine, to design models and problems under study as well as the related experiments to be carried out by our collaborators in biology or medicine. In this context, our ongoing collaborations with biologists and physicians: the collaboration with St Antoine Hospital in Paris within the Institut Universitaire de Cancérologie of UPMC (IUC, Luis Almeida, Jean Clairambault, Dirk Drasdo, Alexander Lorz, Benoît Perthame); Institut Jacques Monod (Luis Almeida); the INRA team headed by Human Rezaei and Wei-Feng Xue's team in the university of Canterbury through the ERC Starting Grant SKIPPER^{AD} (Marie Doumic); our collaborators within the HTE program (François Delhommeau at St Antoine, Thierry Jaffredo, and Delphine Salort at IBPS, UPMC, Paris; François Vallette at INSERM Nantes); Frédéric Thomas at CREEC, Montpellier; Hôpital Paul Brousse through ANR-IFlow and ANR-iLite; and the close experimental collaborations that emerged through the former associate team QUANTISS (Dirk Drasdo), particularly at the Leibniz Institute for Working Environment and Human Factors in Dortmund, Germany, are key points in our project.

Our main objective is the creation, investigation and transfer of new models, methods and algorithms. In selected cases software development as that of CellSys and TiQuant by D. Drasdo and S. Hoehme is performed. More frequently, the team develops “proof of concept” numerical codes in order to test the adequacy of our models to experimental biology.

Taking advantage of the last 4-year evaluation of MAMBA (September 2017), we have re-organised the presentation of our research program in five main axes, three methodological, and two application-driven axes. In more details, these research axes are the following.

Axis 1 (methodological) is devoted to works in physiologically-based design, analysis and control of population dynamics. It encompasses populations of bacteria, of cancer cells, of neurons, of aggregating proteins, etc. whose dynamics are represented by partial differential equations (PDEs), structured in evolving physiological traits, such as cell age, cell size, time elapsed since last firing (neurons).

Axis 2 (methodological) is devoted to reaction and motion equations for living systems. It aims at describing biological phenomena such as tumour growth, chemotaxis and wound healing.

Axis 3 (methodological) tackles the question of model and parameter identification, combining stochastic and deterministic approaches and inverse problem methods in nonlocal and multi-scale models.

Axis 4 (applicative) focuses on cancer, an application on which almost all team members work, with various approaches. A main focus of the team is to study cancer as a Darwinian evolutionary phenomenon in phenotype-structured cell populations. Optimal control methods take into account the two main pitfalls of clinical cancer therapeutics, namely unwanted toxic side effects in healthy cell populations and drug resistance in cancer cell populations. Other studies concern telomere shortening, and multi-scale models.

Axis 5 (applicative) is devoted to growth, evolution and regeneration in populations and tissues. It involves protein aggregation and fragmentation models for neurodegenerative diseases (prion, Alzheimer), organ modelling, mainly of the liver, its damages induced by toxic molecules, and its regeneration after toxic insult. Newcomers in this applicative field are epidemiological modelling of propagation of insect vector-borne diseases by reaction-diffusion equations and of their optimal control, bacterial growth and wound healing.

3.2. Research axis 1: analysis and control for population dynamics

Personnel

Pierre-Alexandre Bliman, Jean Clairambault, Marie Doumic, Alexander Lorz, Benoît Perthame

Project-team positioning

Population dynamics is a field with varied and wide applications, many of them being in the core of MAMBA interests - cancer, bacterial growth, protein aggregation. Their theoretical study also brings a qualitative understanding on the interplay between individual growth, propagation and reproduction in such populations. In the previous periods of evaluation, many results were obtained in the BANG team on the asymptotic and qualitative behaviour of such structured population equations, see e.g. [113], [59], [79], [68]. Other Inria teams interested by this domain are Mycenae, Numed and Dracula, with which we are in close contacts. Among the leaders of the domain abroad, we can cite among others our colleagues Tom Banks (USA), Graeme Wake (New Zealand), Glenn Webb (USA), Jacek Banasiak (South Africa), Odo Diekmann (Netherlands), with whom we are also in regular contact. Most remarkably and recently, connections have also been made with probabilists working on Piecewise Deterministic Markov Processes (F. Malrieu at the university of Rennes, Jean Bertoin at the ETH in Zurich, Vincent Bansaye at Ecole Polytechnique, Julien Berestycki at Cambridge, Amaury Lambert at College de France, M. Hoffmann at Paris Dauphine), leading to a better understanding of the links between both types of results - see also axis 3.

Scientific achievements

We divide this research axis, which relies on the study of structured population equations, according to four different applications, bringing their own mathematical questions, e.g., stability, control, or blow-up.

Time asymptotics for nucleation, growth and division equations

Following the many results obtained in the BANG team on the asymptotic and qualitative behaviour of structured population equation, we put our effort on the investigation of limit cases, where the trend to a steady state or to a steady exponential growth described by the first eigenvector fails to happen. In [65], the case of equal mitosis (division into two equally-sized offspring) with linear growth rate was studied, and strangely enough, it appeared that the general relative entropy method could also be adapted to such a non-dissipative case. Many discussions and common workshops with probabilists, especially through the ANR project PIECE coordinated by F. Malrieu, have led both communities to work closer.

In [77], the case of constant fragmentation rate and linear growth rate has been investigated in a deterministic approach, whereas similar questions were simultaneously raised but in a stochastic process approach in [62].

We also enriched the models by taking into account a nucleation term, modeling the spontaneous formation of large polymers out of monomers [122]. We investigated the interplay between four processes: nucleation, polymerization, depolymerization and fragmentation.

The ERC Starting Grant SKIPPER^{AD} (Domic) supported and was the guideline for the study of nucleation, growth and fragmentation equations.

Cell population dynamics and its control.

One of the important incentives for such model design, source of many theoretical works, is the challenging question of drug-induced drug resistance in cancer cell populations, described in more detail below in axis 4, Cancer. The adaptive dynamics setting used consists of phenotype-structured integro-differential [or reaction-diffusion, when phenotype instability is added under the form of a Laplacian] equations describing the dynamic behaviour of different cell populations interacting in a Lotka-Volterra-like manner that represents

common growth limitation due to scarcity of expansion space and nutrients. The phenotype structure allows us to analyse the evolution in phenotypic traits of the populations under study and its asymptotics for two populations [109], [106], [105], [107]. Space may be added as a complementary structure variable provided that something is known of the (Cartesian) geometry of the population [108], which is seldom the case.

Mathematical models of infectious diseases

These models are made to understand and predict the dynamics of the spread of infectious diseases. We initiated studies with the aim to understand how to use epidemiological data (typically given through incidence rate) in order to estimate the state of the population as well as constants, characteristic of the epidemics such as the transmission rate. The methods rely on observation and identification techniques borrowed from control theory.

Models of neural network

Mean field limits have been proposed by biophysicists in order to describe neural networks based on physiological models. The various resulting equations are called integrate-and-fire, time elapsed models, voltage-conductance models. Their specific nonlinearities and the blow-up phenomena make their originality which has led to develop specific mathematical analysis [116], followed by [112], [101], [117], [67]. This field also yields a beautiful illustration for the capacity of the team to combine and compare stochastic and PDE modelling (see axis 3), in [72].

Collaborations

- Nucleation, growth and fragmentation equations: **Juan Calvo**, university of Granada, came for two one-month visits, **Miguel Escobedo**, University of Bilbao (see also axis 3), **Pierre Gabriel**, University of Versailles-Saint Quentin, former B. Perthame and M. Doumic's Ph.D student, who now co-supervises Hugo Martin's Ph.D thesis.
- Cell population dynamics and its control: **Tommaso Lorenzi**, former Mamba postdoc, now at the University of St. Andrews, Scotland, maintains a vivid collaboration with the Mamba team. He is in particular an external member of the HTE program MoGIIImaging (see also axis 4). **Emmanuel Trélat**, UPMC professor, member of LJLL and of the CAGE Inria team, is the closest Mamba collaborator for optimal control.
- Estimation and identification of epidemiological models: **Maria Soledad Aronna**, Fundação Getulio Vargas, Brazil; **Alain Rapaport**, INRA-Montpellier; **Abderrahmane Iggidr**, Inria Nancy-Grand Est
- Neural networks: **Delphine Salort**, Professor UPMC, Laboratory for computations and quantification in biology, and **Patricia Reynaud**, University of Nice, **Maria Cáceres**, university of Granada.

3.3. Research axis 2: reaction and motion equations for living systems

Personnel

Luis Almeida, Casimir Emako-Kazianou, Alexander Lorz, Benoît Perthame, Nicolas Vauchelet.

Project-team positioning

The Mamba team had initiated and is a leader on the works developed in this research axis. It is a part of a consortium of several mathematicians in France through the ANR Blanc project *Kibord*, which involves in particular members from others Inria team (DRACULA, REO). Finally, we mention that from Sept. 2017 on, Mamba benefits from the ERC Advanced Grant of Benoît Perthame.

Scientific achievements

We divide this research axis, which relies on the study of partial differential equations for space and time organisation of biological populations, according to various applications using the same type of mathematical formalisms and methodologies: asymptotic analysis, weak solutions, numerical algorithms.

Mathematical modelling for bacterial chemotaxis.

Chemotaxis is the phenomenon in which cells direct their motion in response to a chemical signal present in their environment. Our unique expertise is on mathematical aspects of the kinetic equations which describe the run and tumble motion of bacteria and their asymptotic analysis.

An interdisciplinary collaboration with biophysicists from Institut Curie has been successful on experimental observations concerning the interaction between two species of bacteria and emergence of travelling bands [86]. The mathematical models used in this work are derived in [51] thanks to a diffusive limit of a kinetic system with tumbling modulation along the path. A numerical investigation of this limit is provided in [87]. These works enter into the framework of the PhD of Casimir Emako-Kazianou [88]. Recently, we have been able to derive rigorously such kinetic models from a more sophisticated equation incorporating internal variable when cells adapt rapidly to changes in their environment [119].

Aggregation equation.

In the mathematical study of collective behaviour, an important class of models is given by the aggregation equation. In the presence of a non-smooth interaction potential, solutions of such systems may blow up in finite time. To overcome this difficulty, we have defined weak measure-valued solutions in the sense of duality and its equivalence with gradient flows and entropy solutions in one dimension [100]. The extension to higher dimensions has been studied in [70]. An interesting consequence of this approach is the possibility to use the traditional finite volume approach to design numerical schemes able to capture the good behaviour of such weak measure-valued solutions [93], [99].

Free boundary problems for tumour growth.

Fluid dynamic equations are now commonly used to describe tumour growth with two main classes of models: those which describe tumour growth through the dynamics of the density of tumoral cells subjected to a mechanical stress; those describing the tumour through the dynamics of its geometrical domain thanks to a Hele-Shaw-type free boundary model. The first link between these two classes of models has been rigorously obtained thanks to an incompressible limit in [115] for a simple model. This result has motivated the use of another strategy based on viscosity solutions, leading to similar results, in [102].

Since more realistic systems are used in the analysis of medical images, we have extended these studies to include active motion of cells in [114], viscosity in [120] and proved regularity results in [110]. The limiting Hele-Shaw free boundary model has been used to describe mathematically the invasion capacity of a tumour by looking for travelling wave solutions, in [118], see also axis 3. It is a fundamental but difficult issue to explain rigorously the emergence of instabilities in the direction transversal to the wave propagation. For a simplified model, a complete explanation is obtained in [103].

Collaborations

- Institut Curie, joint work with Axel Buguin on bacterial models for chemotaxis.
- Shanghai Jiao Tong University, joint publications with Min Tang on bacterial models for chemotaxis and free boundary problems for tumour growth.
- Imperial College London, joint works with José Antonio Carrillo on aggregation equation.
- University of Maryland at College Park, UCLA, Univ. of Chicago, Univ. Autónoma de Madrid, Univ. of St. Andrews (Scotland), joint works on mathematics of tumour growth models.

3.4. Research axis 3: Model and parameter identification combining stochastic and deterministic approaches in nonlocal and multi-scale models

Personnel

Marie Doumic, Dirk Drasdo, Aurora Armiento, Thibault Bourgeron, Rebecca Chisholm, Tommaso Lorenzi

Project-team positioning

Mamba developed and addressed model and parameter identification methods and strategies in a number of mathematical and computational model applications including growth and fragmentation processes emerging in bacterial growth and protein misfolding, in liver regeneration [82], TRAIL treatment of HeLa cells [61], growth of multicellular spheroids [98], blood detoxification after drug-induced liver damage [124], [92].

This naturally led to increasingly combine methods from various fields: image analysis, statistics, probability, numerical analysis, PDEs, ODEs, agent-based modelling methods, involving inverse methods as well as direct model and model parameter identification in biological and biomedical applications. Model types comprise agent-based simulations for which Mamba is among the leading international groups, and Pharmacokinetic (PK) simulations that have recently combined in integrated models (PhD theses Géraldine Cellière, Noémie Boissier). The challenges related with the methodological variability has led to very fruitful collaborations with internationally renowned specialists of these fields, e.g. for bacterial growth and protein misfolding with Marc Hoffmann (Paris Dauphine) and Patricia Reynaud-Bouret (University of Nice) in statistics, with Philippe Robert (Inria RAP) in probability, with Tom Banks (Raleigh, USA) and Philippe Moireau (Inria M3DISIM) in inverse problems and data assimilation, and with numerous experimentalists.

Scientific achievements

Direct parameter identification is a great challenge particularly in living systems in which part of parameters at a certain level are under control of processes at smaller scales.

Estimation methods for growing and dividing populations

In this domain, all originated in two papers in collaboration with J.P. Zubelli in 2007 [121], [81], whose central idea was to use the asymptotic steady distribution of the individuals to estimate the division rate. A series of papers improved and extended these first results while keeping the deterministic viewpoint, lastly [65]. The last developments now tackle the still more involved problem of estimating not only the division rate but also the fragmentation kernel (i.e., how the sizes of the offspring are related to the size of the dividing individual) [35]. In parallel, in a long-run collaboration with statisticians, we studied the Piecewise Deterministic Markov Process (PDMP) underlying the equation, and estimated the division rate directly on sample observations of the process, thus making a bridge between the PDE and the PDMP approach in [80], a work which inspired also very recently other groups in statistics and probability [62], [95] and was the basis for Adélaïde Olivier's Ph.D thesis [111], [96] and of some of her more recent works [21] (see also axis 5).

Model identification for growing multicellular spheroids

For multicellular spheroids growing under different conditions, first an agent-based model on an unstructured lattice for one condition has been developed and then stepwise extended for each additional condition which could not be captured by the present model state [98] [97] (axis 5). The multicellular dynamics has been mimicked by a master equation, intracellular processes by ODEs, and extracellular molecular transport processes by partial differential equations. The model development was based almost completely on bright field image sequences, whereby image segmentation parameter identification was performed by investigation of sensitivity and specificity of the segmentation with a biologist expert serving as gold standard. The Akaike and Bayesian Information Criteria were used to evaluate whether parameters introduced due to the model extension led to a significant increase of the information. It turned out that the final model could predict the outcome of growth conditions not considered in model development.

A similar stepwise strategy has been recently performed to identify the pressure and strain constraints of growing multicellular cell populations subject to mechanical stress. Here, a novel deformable cell model has been used that permits to display cell shape changes explicitly [127].

Data assimilation and stochastic modelling for protein aggregation

Estimating reaction rates and size distributions of protein polymers is an important step for understanding the mechanisms of protein misfolding and aggregation (see also axis 5). In [52], we settled a framework problem when the experimental measurements consist in the time-dynamics of a moment of the population.

To model the intrinsic variability among experimental curves in aggregation kinetics - an important and poorly understood phenomenon - Sarah Eugène's Ph.D, co-supervised by P. Robert [89], was devoted to the stochastic modelling and analysis of protein aggregation, compared both with the deterministic approach traditionally developed in Mamba [122] and with experiments.

Model identification in liver regeneration

Based on successful model predictions for models addressing different aspects of liver regeneration, we extracted a general workflow on how modelling can inform liver disease pathogenesis [82]. Liver has a complex micro-architecture ensuring its function (axis 5). Hence many clinical questions require quantitative characterisation of micro-architecture in normal liver and during degeneration or regeneration processes which has been performed using the software TiQuant ([91] see software) or other tools we generated. Disease specific and personal information can be used to build a list of hypotheses on the question of interest, which then can be systematically implemented in mathematical models and in simulation runs tested against the data. As confocal micrographs only display part of lobules, statistically representative lobules were constructed to permit definition of boundary conditions for flow and transport.

In order to compare data and model results quantitatively, quantitative measures characterising the processes under study have to be defined, and measured in both experiment and model. Models in a context where micro-architecture is important were based on agent-based models representing each hepatocyte as well as blood vessels explicitly. They were parameterised by measurable parameters as for those physiologically relevant ranges can be identified, and systematic simulated parameter sensitivity analyses can be performed. Movement of each cell was then mimicked by an equation of motion, describing the change of position as a function of all forces on that cell including active migration. If the best model disagreed with the data, the underlying hypotheses were considered incomplete or wrong and were modified or complemented. Models quantitatively reproducing data were either used to predict so far unknown situations, or the key mechanisms were directly challenged by our experimental partners. Along this line, two unrecognised mechanisms could be identified (axis 5).

Statistical methods decide on subsequently validated mechanism of ammonia detoxification

To identify the mechanisms involved in ammonia detoxification [92], 8 candidate models representing the combination of three possible mechanisms were developed (axis 5). First, the ability of each model to capture the experimental data was assessed by statistically testing the null hypothesis that the data have been generated by the model, leading to exclusion of one of the 8 models. The 7 remaining models were compared among each other by the likelihood ratio. The by far best models were those containing a particular ammonia sink mechanism, later validated experimentally (axis 5). For each of the statistical tests, the corresponding test statistics has been calculated empirically and turned out to be not chi2-distributed in opposition to the usual assumption stressing the importance of calculating the empirical distribution, especially when some parameters are unidentifiable.

Collaborations

- **Philippe Robert**, Inria Rap, for the stochastic process modelling [90]
- **Marc Hoffmann**, Université Paris-Dauphine, for the statistical approach to growth and division processes [80], **M. Escobedo**, Bilbao and **M. Tournus**, Marseille, for the deterministic approach.
- **Tom Banks**, North Carolina State University, and **Philippe Moireau**, Inria M3DISIM, for the inverse problem and data assimilation aspects [57],[1]
- **Jan G. Henstler group**, IfADo, Dortmund (Germany), **Irene Vignon-Clementel** (Inria, REO), others for Liver regeneration, ammonia detoxification.
- **Kai Breuhahn group**, DKFZ Heidelberg (Germany), **Pierre Nassoy**, Univ. Bordeaux, for multicellular tumor growth

3.5. Research axis 4: Focus on cancer

Personnel

Luis Almeida, Thibault Bourgeron, Cécile Carrère, Rebecca Chisholm, Jean Clairambault, Marie Doumic, Dirk Drasdo, Sarah Eugène, Paul Van Liedekerke, Tommaso Lorenzi, Alexander Lorz, Benoît Perthame, Yi Yin

Project-team positioning

The MAMBA team designs and analyses mathematical models of tumour growth and therapy, at the cell population level, using agent-based or partial differential equations, with special interest in methodologies for therapeutic optimisation using combined anticancer drug treatments. Rather than, or not only, modelling the effect of drugs on molecular targets, we represent these effects by their *functional* consequences on the fate of healthy and cancer cell populations: proliferation (velocity of the cell division cycle, decreasing it, e.g., by antagonising growth factor receptors), apoptosis, cell death or senescence.

Our goal in doing this is to circumvent the two main issues of anticancer therapy in the clinic, namely unwanted toxic side effects in populations of healthy cells and emergence of drug-induced drug resistance in cancer cell populations. This point of view leads us to take into account phenomena of transient and reversible resistance, observed in many cancer cell populations, by designing and analysing models of cell populations structured in continuous phenotypes, relevant for the description of the behaviour of cell populations exposed to drugs: either degree of resistance to a given drug, or potential of resistance to drug-induced stress, proliferation potential, and plasticity.

Such modelling options naturally lead us to take into account in a continuous way (i.e., by continuous-valued phenotype or relevant gene expression) the wide phenotypic heterogeneity of cancer cell populations. They also lead us to adopt the point of view of *adaptive dynamics* according to which characteristic traits of cell populations evolve with tumour environmental pressure (drugs, cytokines or metabolic conditions, mechanical stress and spatial conditions), in particular from drug sensitivity to resistance. This position is original on the international scene of teams dealing with drug resistance in cancer.

Scientific achievements

Molecular modelling towards theoretical optimisation of anticancer drug delivery

The protein p53, guardian of the genome and tumour suppressor, has been the object of Ján Eliaš's PhD thesis [85], defended in September 2015, and of articles in 2014 and 2017 [83], [84], [11]. Based on an original intracellular spatial PDE model of the protein dynamics, it allows for the prediction of biologically observed oscillations of p53 nuclear concentrations in case of (e.g. radiotherapy- or anticancer drug-induced) damage to the DNA. In parallel, in [75], that for us concluded works initiated by a fruitful collaboration with Francis Lévi (retired from CNRS 2014), we associate pharmacokinetics-pharmacodynamics of anticancer drugs, their action on the cell cycle at the cell population level, and optimisation algorithms to maximise their combined action under the constraint of preserving healthy tissue integrity.

Modelling Acute Myeloid Leukaemia (AML) and its control by anticancer drugs by PDEs and Delay Differential equations

In collaboration with Catherine Bonnet (Inria DISCO, Saclay) and François Delhommeau (St Antoine hospital in Paris), together with DISCO PhD students José Luis Avila Alonso and Walid Djema, this theme has led to common published proceedings of conferences: IFAC, ACC, CDC, MTNS [54], [55], [56], [64], [76], [53]. These works study the stability of the haematopoietic system and its possible restabilisation by combinations of anticancer drugs with functional targets on cell populations: proliferation, apoptosis, differentiation.

Adaptive dynamics setting to model and circumvent evolution towards drug resistance in cancer by optimal control

We tackle the problem to represent and inhibit - using optimal control algorithms, in collaboration with Emmanuel Trélat, proposed Inria team CAGE - drug-induced drug resistance in cancer cell populations. This theme, presently at the core of our works on cancer modelling with a evolutionary perspective on tumour heterogeneity, is documented in a series of articles [73], [74], [105], [106], [108]. Taking into account the two main pitfalls of cancer therapy, unwanted side effects on healthy cells and evolution towards resistance in cancer cells, it has attracted to our team the interest of several teams of biologists, with whom we have undertaken common collaborative works, funded by laureate answers to national calls (see ITMO Cancer HTE call).

This theme is also at the origin of methodological developments (see Research axis 1)

Senescence modelling by telomere shortening

In many animals, aging tissues accumulate senescent cells, a process which is beneficial to protect from cancer in the young organism. In collaboration with Teresa Teixeira and Zhou Xu from IBCP, we proposed a mathematical model based on the molecular mechanisms of telomere replication and shortening and fitted it on individual lineages of senescent *Saccharomyces cerevisiae* cells, in order to decipher the causes of heterogeneity in replicative senescence [66].

Biomechanically mediated growth control of cancer cells

Mechanical feedback has been identified as a key regulator of tissue growth, by which external signals are transduced into a complex intracellular molecular machinery. Using multi-scale computational modelling of multicellular growth in two largely different experimental settings with the same tumour cell line we were able to show that the cellular growth response on external mechanical stress is surprisingly quantitatively predictable. For this purpose, the mechanical parameters of a center-based agent-based model were calibrated with a deformable agent-based cell model, which displays cell shape and hence can deal with high cell compressions. The cell cycle progression function was calibrated with findings of population growth in an elastic capsule. The emerging model was able to correctly predict the growth response both for modified stresses in a capsule as well as the growth response in a different experimental setting [128], [127].

Model identification for TRAIL treatment

Repetitive administration of TRAIL (TNF-Related Apoptosis Induced-Ligand) on HeLa cells produces characteristic resistance pattern in time that can be explained by cell-to-cell variability in the protein composition. The TRAIL signal transduction pathway is one of the best-studied apoptosis pathways and hence permits detailed comparisons with data. Within a stochastic model of gene expression coupled to transcription and translation to the pathway members, we were able to quantitatively explain the resistance pattern. An important challenge was in parameter identification at each of the level for numerous proteins, whereby the most sensitive parameter was to correctly capture short-lived proteins in the TRAIL toxicity pathway as those mainly determine the regeneration of protein distribution in the cell population and thereby may generate strong stochastic fluctuations [61], [60].

Radiotherapy

In close cooperation with M. Herrero (U. Complutense, Madrid) we have explored by extensive computer simulations using an agent-based model the consequences of spatially inhomogeneous x-ray irradiation in cancer treatment. The model predicted that in the case of different competing sub-populations, namely cancer stem cells with unlimited division capacity, and cancer cells with limited division capacity, inhomogeneous radiation focusing higher doses at the tumour center and lower doses at the tumour periphery should outperform homogeneous irradiation [104]. Cancer stem cells are believed to have a longer cell cycle duration than cancer cells, and are less radiosensitive than cancer cells, which is why they often survive radiation and lead to tumour relapse.

Collaborations

- AML modelling: **Catherine Bonnet**, DISCO Inria team, Saclay, and **François Delhommeau**, INSERM St Antoine (also collaborator in the INSERM HTE laureate project EcoAML, see below).
- INSERM HTE laureate project MoGImaging, headed by E. Moyal (Toulouse): **François Vallette**, CRCNA and INSERM Nantes
- INSERM HTE laureate project EcoAML, headed by **François Delhommeau**, INSERM St Antoine: François Delhommeau, Thierry Jaffredo (IBPS), Delphine Salort (LCQB-IBPS)
- Adaptive dynamics to model drug resistance and optimal control to circumvent it:

Alexandre Escargueil (2 articles in common [74], [108]), **Michèle Sabbah** (2 PhD theses in common) at Annette Larsen's lab, St Antoine hospital, Paris

Emmanuel Trélat (1 PhD thesis in common) at Inria team CAGE and Laboratoire Jacques-Louis Lions at Sorbonne Université.

Frédéric Thomas at CREEC, Montpellier: one funded Inria PRE project in common.

- Telomere shortening: **Teresa Teixeira** and **Zhou Xu** (IBCP, Paris), **Philippe Robert** (Inria RAP).
- TRAIL treatment: **Gregory Batt**, Inria Saclay and Inst. Pasteur (France)

3.6. Research axis 5: Growth, evolution and regeneration in populations and tissues

Personnel

Luis Almeida, Pierre-Alexandre Bliman, Marie Doumic, Dirk Drasdo, Benoît Perthame, Nicolas Vauchelet

Project-team positioning

The applications in this category span very different subjects from amyloid diseases, dengue fever, wound healing, liver regeneration and toxicity, up to bacterial growth. As the applications, the methods span a wide range. Those concerning identification of models and parameters with regard to data have partially been outlined in axis 3. Focus in this axis is on the model contribution to the biologically and/or medically relevant insights and aspects.

Liver-related modelling is partially performed within the Inria team MIMESIS (Strasbourg) with the focus on real-time, patient-specific biomechanical liver models to guide surgery and surgeons. Internationally, spatial temporal liver related models are developed in Fraunhofer MEVIS (Bremen), by T. Ricken (TU Dortmund), and P. Segers group (Leuven). Different from these, Mamba has a strong focus on spatial-temporal modelling on the histological scale, integration of molecular processes in each individual cell, and single-cell (agent) based models. Works by Schliess [124], [92] have been highlighted in editorials.

Mathematical modelling of protein aggregation is a relatively recent domain, only a few other groups have emerged yet; among them we can cite the Inria team Dracula, with whom we are in close contact, and e.g., the work by Jean-Michel Coron (UPMC) and Monique Chyba (Hawaii, USA) in control, and Suzanne Sindi (USA) for the modelling of the yeast prion. We have interactions with all these groups and organised a workshop in June 2017, gathering both the biophysics and applied mathematics communities.

Scientific achievements

Amyloid disease

Application to protein aggregation in amyloid diseases is a long-standing interest of Mamba, dating back to 2010 [69], and developed through the collaboration with Human Rezaei's team at Inra. More recently, with Wei-Feng Xue in Canterbury, we investigated the intrinsic variability among identical experiments of nucleation [78], [90], Sarah Eugène's Ph.D subject (co-supervised by Philippe Robert) [89].

In collaboration with Tom Banks first [58], [57] and then Philippe Moireau, we developed quantitative comparisons between model and data. Through data assimilation and statistical methods [52], we proposed new models and mechanisms.

Dengue fever

The spread of certain strains of the intracellular parasitic bacterium *Wolbachia* in populations of mosquitoes *Aedes aegypti* drastically reduces their competence as vector of dengue and other severe mosquito-borne viral diseases. In the absence of a vaccine, or of any preventive or curative treatment, the release of mosquitoes deliberately infected in laboratory by this bacterium has been recently considered a promising tool to control these diseases. Technically the situation can be described by a bistable model, and the issue consists in moving from a *Wolbachia*-free equilibrium to a fully contaminated equilibrium.

When implementing such a method, an important issue concerns the spatial propagation of the mosquitoes: on releasing infected mosquitoes in a given domain (which can be part of a city), the hope is to invade the whole area. The study of this propagation phenomena falls into the study of existence of travelling waves.

Wound healing

We studied cell motion in epithelial gap closure, a form of collective cell migration that is a very widespread phenomenon both during development and adult life - it is essential for both the formation and for the maintenance of epithelial layers. Due to their importance, *in vivo* wound healing and morphogenetic movements involving closure of holes in epithelia have been the object of many studies. In our works⁰ we considered wound healing and epithelial gap closure in both *in vivo* (in particular drosophila pupa) and *in vitro* (MDCK cell and human keratinocytes). We found some similarities in the geometry dependence of the wound closure strategies between these two situations, indicating the existence of conserved mechanisms that should be widespread across living beings.

Liver regeneration

An integrated model, coupling a spatial-temporal model of liver regeneration after drug-induced damage to a compartment model of detoxification blood from ammonia, identified the lack of an ammonia detoxifying reaction in the biochemical consensus scheme [124]. Hyperammonia is the most frequent reason for death due to acute liver failure in UK and USA. The spatial model represents liver micro-architecture in a group of liver lobules, the repetitive anatomical and functional units of liver, mimicking each hepatocyte as single agent and blood vessels as a network of chains of spherical objects. This model had previously predicted the subsequently validated orientation of dividing hepatocytes along the liver capillaries as order mechanism. It was here coupled to ODEs for metabolites participating in the zoned ammonia metabolism by calculating the volume of each liver lobule zone with time during regeneration after drug induced damage, which is an input parameter for the detoxification compartment model. Experiments triggered by the model predictions could identify later a candidate ammonia sink mechanism which in a follow-up work [92] could be shown to be the most likely mechanism compared with alternative explanations (see axis 3). This mechanism could be validated, and led to a possible therapy option in treatment of hyperammonemia.

The models have been further expanded towards true multilevel-multiscale models that include molecular HGF control of cell cycle progression (unpublished) and ammonia detoxification (Géraldine Cellière's PhD thesis, 2016; Noémie Boissier's PhD thesis, 2018). In these models, the intracellular models were executed in each individual hepatocyte, and transport of molecules with blood were simulated. Blood flow was modelled by Poiseuille law in the entire capillary network. Further conditions could be identified, under which standard pharmacokinetics-pharmacodynamics (PKPD) models fail to predict the correct dynamics and need to be replaced by spatial temporal models representing organ microarchitecture. The model has further been extended towards bile flow.

Toxicity extrapolation from *in vitro* to *in vivo*

In vivo toxicity prediction from *in vitro* data is a major objective in toxicology as it permits bypassing animal experiments. The multilevel-multi scale approach outlined above has been used to explore a strategy to predict the *in vivo* damage of paracetamol (acetaminophen) from *in vitro* experimental data. Model simulations and data obtained so far strongly suggest that the prediction is quantitative, if the time development of the toxicity *in vitro* is displayed (this is so far not common), differences in the concentration kinetics of drug metabolising enzymes *in vitro* are measured, and micro-architecture is determined (Géraldine Cellière's PhD thesis [71]). Common strategies in toxicology based on relating the maximum drug concentration or area under the drug concentration - time curve between *in vitro* and *in vivo* damage could be shown to fail.

Bacterial population growth

We exploited all the methods developed to estimate the division rate of a population (see axis 3) to address a seminal question of biology: is it a size-sensing or a timing mechanism which triggers bacterial growth? In [123], we showed that a sizer model is robust and fits the data well. Several studies from other groups came at the same time, showing a renewed interest on a question dated back to Jacques Monod's PhD thesis (1941). Of special interest is the "adder" model, for which we are currently developing new estimation methods.

⁰ravasio:hal-01245750, vedula:hal-01298859

Collaborations

- Dengue control by releasing Wolbachia infected mosquitoes **Maria Soleda Aronna**, **F.C. Coelho** (Fundação Getulio Vargas, Brazil); **D. Villela**, **C. Struchiner** (Fiocruz, Brazil); **Jorge Zubelli** (IMPA, Brazil); **Alain Rapaport** (INRA-Montpellier), **Y. Dumont** (CIRAD-Montpellier); **Ch. Schaerer** (UNA, Paraguay).
- Protein aggregation in amyloid diseases: **Human Rezaei**'s team at Inra Jouy-en-Josas (France) and **W-F Xue**'s team in at university of Kent (Great Britain); **Tom Banks** at the North Carolina State University (USA), **Philippe Moireau** (M3DISIM) and **Philippe Robert** (Rap) in Inria
- bacterial growth and division: **Lydia Robert**, UPMC (France)
- Liver research & toxicology: **JG. Hengstler** group (IfADo, Dortmund, Germany); **R. Gebhardt** (Univ. Leipzig); **U. Klingmueller** (DKFZ, Heidelberg); **Irène Vignon-Clementel** (Inria, REO)
- Wound healing: **Patrizia Bagnerini** (Genova, Numerical methods), **Benoît Ladoux** (Institut Jacques Monod et Mechanobiology Institute Singapore, Biophysics) and **Antonio Jacinto** (CEDOC, Lisbon, Biology and Medicine).

MONC Project-Team

3. Research Program

3.1. Introduction

We are working in the context of data-driven medicine against cancer. We aim at coupling mathematical models with data to address relevant challenges for biologists and clinicians in order for instance to improve our understanding in cancer biology and pharmacology, assist the development of novel therapeutic approaches or develop personalized decision-helping tools for monitoring the disease and evaluating therapies.

More precisely, our research on mathematical oncology is three-fold:

- Axis 1: Tumor modeling for patient-specific simulations: *Clinical monitoring. Numerical markers from imaging data. Radiomics.*
- Axis 2: Bio-physical modeling for personalized therapies: *Electroporation from cells to tissue. Radiotherapy.*
- Axis 3: Quantitative cancer modeling for biological and preclinical studies: *Biological mechanisms. Metastatic dissemination. Pharmacometrics.*

In the first axis, we aim at producing patient-specific simulations of the growth of a tumor or its response to treatment starting from a series of images. We hope to be able to offer a valuable insight on the disease to the clinicians in order to improve the decision process. This would be particularly useful in the cases of relapses or for metastatic diseases.

The second axis aims at modeling biophysical therapies like radiotherapies, but also thermo-ablations, radio-frequency ablations or electroporation that play a crucial role in the case of a relapse or for a metastatic disease, which is precisely the clinical context where the techniques of axis 1 will be applied.

The third axis, even if not directly linked to clinical perspectives, is essential since it is a way to better understand and model the biological reality of cancer growth and the (possibly complex) effects of therapeutic intervention. Modeling in this case also helps to interpret the experimental results and improve the accuracy of the models used in Axis 1. Technically speaking, some of the computing tools are similar to those of Axis 1.

3.2. Axis 1: Tumor modeling for patient-specific simulations

The gold standard treatment for most cancers is surgery. In the case where total resection of the tumor is possible, the patient often benefits from an adjuvant therapy (radiotherapy, chemotherapy, targeted therapy or a combination of them) in order to eliminate the potentially remaining cells that may not be visible. In this case personalized modeling of tumor growth is useless and statistical modeling will be able to quantify the risk of relapse, the mean progression-free survival time...However if total resection is not possible or if metastases emerge from distant sites, clinicians will try to control the disease for as long as possible. A wide set of tools are available. Clinicians may treat the disease by physical interventions (radiofrequency ablation, cryoablation, radiotherapy, electroporation, focalized ultrasound,...) or chemical agents (chemotherapies, targeted therapies, antiangiogenic drugs, immunotherapies, hormonotherapies). One can also decide to monitor the patient without any treatment (this is the case for slowly growing tumors like some metastases to the lung, some lymphomas or for some low grade glioma). A reliable patient-specific model of tumor evolution with or without therapy may have different uses:

- Case without treatment: the evaluation of the growth of the tumor would offer a useful indication for the time at which the tumor will reach a critical size. For example, radiofrequency ablation of pulmonary lesion is very efficient as long as the diameter of the lesion is smaller than 3 cm. Thus, the prediction can help the clinician plan the intervention. For slowly growing tumors, quantitative modeling can also help to decide at what time interval the patient has to undergo a CT-scan. CT-scans

are irradiative exams and there is a challenge for decreasing their occurrence for each patient. It has also an economical impact. And if the disease evolution starts to differ from the prediction, this might mean that some events have occurred at the biological level. For instance, it could be the rise of an aggressive phenotype or cells that leave a dormancy state. This kind of events cannot be predicted, but some mismatch with respect to the prediction can be an indirect proof of their existence. It could be an indication for the clinician to start a treatment.

- Case with treatment: a model can help to understand and to quantify the final outcome of a treatment using the early response. It can help for a redefinition of the treatment planning. Modeling can also help to anticipate the relapse by analyzing some functional aspects of the tumor. Again, a deviation with respect to reference curves can mean a lack of efficiency of the therapy or a relapse. Moreover, for a long time, the response to a treatment has been quantified by the RECIST criteria which consists in (roughly speaking) measuring the diameters of the largest tumor of the patient, as it is seen on a CT-scan. This criteria is still widely used and was quite efficient for chemotherapies and radiotherapies that induce a decrease of the size of the lesion. However, with the systematic use of targeted therapies and anti-angiogenic drugs that modify the physiology of the tumor, the size may remain unchanged even if the drug is efficient and deeply modifies the tumor behavior. One better way to estimate this effect could be to use functional imaging (Pet-scan, perfusion or diffusion MRI, ...), a model can then be used to exploit the data and to understand in what extent the therapy is efficient.
- Optimization: currently, we do not believe that we can optimize a particular treatment in terms of distribution of doses, number, planning with the model that we will develop in a medium term perspective. But it is an aspect that we keep in mind on a long term one.

The scientific challenge is therefore as follows: knowing the history of the patient, the nature of the primitive tumor, its histopathology, knowing the treatments that patients have undergone, knowing some biological facts on the tumor and having a sequence of images (CT-scan, MRI, PET or a mix of them), are we able to provide a numerical simulation of the extension of the tumor and of its metabolism that fits as best as possible with the data (CT-scans or functional data) and that is predictive in order to address the clinical cases described above?

Our approach relies on the elaboration of PDE models and their parametrization with the image by coupling deterministic and stochastic methods. The PDE models rely on the description of the dynamics of cell populations. The number of populations depends on the pathology. For example, for glioblastoma, one needs to use proliferative cells, invasive cells, quiescent cells as well as necrotic tissues to be able to reproduce realistic behaviors of the disease. In order to describe the relapse for hepatic metastases of gastro-intestinal stromal tumor (gist), one needs three cell populations: proliferative cells, healthy tissue and necrotic tissue.

The law of proliferation is often coupled with a model for the angiogenesis. However such models of angiogenesis involve too many non measurable parameters to be used with real clinical data and therefore one has to use simplified or even simplistic versions. The law of proliferation often mimics the existence of an hypoxia threshold, it consists of an ODE. or a PDE that describes the evolution of the growth rate as a combination of sigmoid functions of nutrients or roughly speaking oxygen concentration. Usually, several laws are available for a given pathology since at this level, there are no quantitative argument to choose a particular one.

The velocity of the tumor growth differs depending on the nature of the tumor. For metastases, we will derive the velocity thanks to Darcy's law in order to express that the extension of the tumor is basically due to the increase of volume. This gives a sharp interface between the metastasis and the surrounding healthy tissues, as observed by anatomopathologists. For primitive tumors like glioma or lung cancer, we use reaction-diffusion equations in order to describe the invasive aspects of such primitive tumors.

The modeling of the drugs depends on the nature of the drug: for chemotherapies, a death term can be added into the equations of the population of cells, while antiangiogenic drugs have to be introduced in an angiogenic model. Resistance to treatment can be described either by several populations of cells or with non-constant growth or death rates. As said before, it is still currently difficult to model the changes of phenotype

or mutations, we therefore propose to investigate this kind of phenomena by looking at deviations of the numerical simulations compared to the medical observations.

The calibration of the model is achieved by using a series (at least 2) of images of the same patient and by minimizing a cost function. The cost function contains at least the difference between the volume of the tumor that is measured on the images with the computed one. It also contains elements on the geometry, on the necrosis and any information that can be obtained through the medical images. We will pay special attention to functional imaging (PET, perfusion and diffusion MRI). The inverse problem is solved using a gradient method coupled with some Monte-Carlo type algorithm. If a large number of similar cases is available, one can imagine to use statistical algorithms like random forests to use some non quantitative data like the gender, the age, the origin of the primitive tumor...for example for choosing the model for the growth rate for a patient using this population knowledge (and then to fully adapt the model to the patient by calibrating this particular model on patient data) or for having a better initial estimation of the modeling parameters. We have obtained several preliminary results concerning lung metastases including treatments and for metastases to the liver.

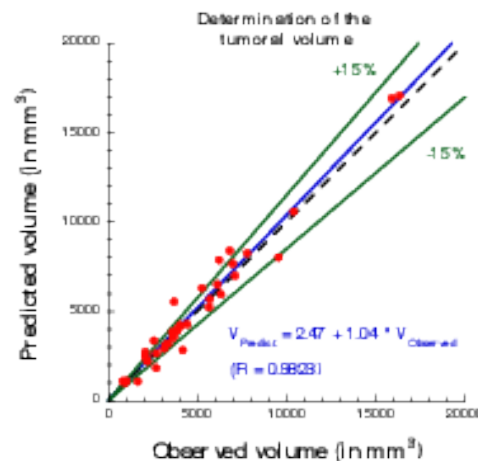


Figure 4. Plot showing the accuracy of our prediction on meningioma volume. Each point corresponds to a patient whose two first exams were used to calibrate our model. A patient-specific prediction was made with this calibrated model and compared with the actual volume as measured on a third time by clinicians. A perfect prediction would be on the black dashed line. Medical data was obtained from Prof. Loiseau, CHU Pellegrin.

3.3. Axis 2: Bio-physical modeling for personalized therapies

In this axis, we investigate locoregional therapies such as radiotherapy, irreversible electroporation. Electroporation consists in increasing the membrane permeability of cells by the delivery of high voltage pulses. This non-thermal phenomenon can be transient (reversible) or irreversible (IRE). IRE or electro-chemotherapy – which is a combination of reversible electroporation with a cytotoxic drug – are essential tools for the treatment of a metastatic disease. Numerical modeling of these therapies is a clear scientific challenge. Clinical applications of the modeling are the main target, which thus drives the scientific approach, even though theoretical studies in order to improve the knowledge of the biological phenomena, in particular for electroporation, should also be addressed. However, this subject is quite wide and we focus on two particular approaches: some aspects of radiotherapies and electro-chemotherapy. This choice is motivated partly by pragmatic reasons: we

already have collaborations with physicians on these therapies. Other treatments could be probably treated with the same approach, but we do not plan to work on this subject on a medium term.

- Radiotherapy (RT) is a common therapy for cancer. Typically, using a CT scan of the patient with the structures of interest (tumor, organs at risk) delineated, the clinicians optimize the dose delivery to treat the tumor while preserving healthy tissues. The RT is then delivered every day using low resolution scans (CBCT) to position the beams. Under treatment the patient may lose weight and the tumor shrinks. These changes may affect the propagation of the beams and subsequently change the dose that is effectively delivered. It could be harmful for the patient especially if sensitive organs are concerned. In such cases, a replanification of the RT could be done to adjust the therapeutical protocol. Unfortunately, this process takes too much time to be performed routinely. The challenges faced by clinicians are numerous, we focus on two of them:
 - *Detecting the need of replanification:* we are using the positioning scans to evaluate the movement and deformation of the various structures of interest. Thus we can detect whether or not a structure has moved out of the safe margins (fixed by clinicians) and thus if a replanification may be necessary. In a retrospective study, our work can also be used to determine RT margins when there are no standard ones. A collaboration with the RT department of Institut Bergonié is underway on the treatment of retroperitoneal sarcoma and ENT tumors (head and neck cancers). A retrospective study was performed on 11 patients with retro-peritoneal sarcoma. The results have shown that the safety margins (on the RT) that clinicians are currently using are probably not large enough. The tool used in this study was developed by an engineer funded by Inria (Cynthia Périer, ADT Sesar). We used well validated methods from a level-set approach and segmentation / registration methods. The originality and difficulty lie in the fact that we are dealing with real data in a clinical setup. Clinicians have currently no way to perform complex measurements with their clinical tools. This prevents them from investigating the replanification. Our work and the tools developed pave the way for easier studies on evaluation of RT plans in collaboration with Institut Bergonié. *There was no modeling involved in this work that arose during discussions with our collaborators.* The main purpose of the team is to have meaningful outcomes of our research for clinicians, sometimes it implies leaving a bit our area of expertise.
 - *Evaluating RT efficacy and finding correlation between the radiological responses and the clinical outcome:* our goal is to help doctors to identify correlation between the response to RT (as seen on images) and the longer term clinical outcome of the patient. Typically, we aim at helping them to decide when to plan the next exam after the RT. For patients whose response has been linked to worse prognosis, this exam would have to be planned earlier. This is the subject of collaborations with Institut Bergonié and CHU Bordeaux on different cancers (head and neck, pancreas). The response is evaluated from image markers (e.g. using texture information) or with a mathematical model developed in Axis 1. The other challenges are either out of reach or not in the domain of expertise of the team. Yet our works may tackle some important issues for adaptive radiotherapy.
- Both IRE and electrochemotherapy are anticancerous treatments based on the same phenomenon: the electroporation of cell membranes. This phenomenon is known for a few decades but it is still not well understood, therefore our interest is two fold:
 1. We want to use mathematical models in order to better understand the biological behavior and the effect of the treatment. We work in tight collaboration with biologists and bioelectromagneticians to derive precise models of cell and tissue electroporation, in the continuity of the research program of the Inria team-project MC2. These studies lead to complex non-linear mathematical models involving some parameters (as less as possible). Numerical methods to compute precisely such models and the calibration of the parameters with the experimental data are then addressed. Tight collaborations with the Vectorology and Anticancerous Therapies (VAT) of IGR at Villejuif, Laboratoire Ampère of Ecole

Centrale Lyon and the Karlsruhe Institute of technology will continue, and we aim at developing new collaborations with Institute of Pharmacology and Structural Biology (IPBS) of Toulouse and the Laboratory of Molecular Pathology and Experimental Oncology (LM-PEO) at CNR Rome, in order to understand differences of the electroporation of healthy cells and cancer cells in spheroids and tissues.

2. This basic research aims at providing new understanding of electroporation, however it is necessary to address, particular questions raised by radio-oncologists that apply such treatments. One crucial question is "What pulse or what train of pulses should I apply to electroporate the tumor if the electrodes are located as given by the medical images"? Even if the real-time optimization of the placement of the electrodes for deep tumors may seem quite utopian since the clinicians face too many medical constraints that cannot be taken into account (like the position of some organs, arteries, nerves...), one can expect to produce real-time information of the validity of the placement done by the clinician. Indeed, once the placement is performed by the radiologists, medical images are usually used to visualize the localization of the electrodes. Using these medical data, a crucial goal is to provide a tool in order to compute in real-time and visualize the electric field and the electroporated region directly on these medical images, to give the doctors a precise knowledge of the region affected by the electric field. In the long run, this research will benefit from the knowledge of the theoretical electroporation modeling, but it seems important to use the current knowledge of tissue electroporation – even quite rough –, in order to rapidly address the specific difficulty of such a goal (real-time computing of non-linear model, image segmentation and visualization). Tight collaborations with CHU Pellegrin at Bordeaux, and CHU J. Verdier at Bondy are crucial.
- Radiofrequency ablation. In a collaboration with Hopital Haut Leveque, CHU Bordeaux we are trying to determine the efficacy and risk of relapse of hepatocellular carcinoma treated by radiofrequency ablation. For this matter we are using geometrical measurements on images (margins of the RFA, distance to the boundary of the organ) as well as texture information to statistically evaluate the clinical outcome of patients.
 - Intensity focused ultrasound. In collaboration with Utrecht Medical center, we aim at tackling several challenges in clinical applications of IFU: target tracking, dose delivery...

3.4. Axis 3: Quantitative cancer modeling for biological and preclinical studies

With the emergence and improvement of a plethora of experimental techniques, the molecular, cellular and tissue biology has operated a shift toward a more quantitative science, in particular in the domain of cancer biology. These quantitative assays generate a large amount of data that call for theoretical formalism in order to better understand and predict the complex phenomena involved. Indeed, due to the huge complexity underlying the development of a cancer disease that involves multiple scales (from the genetic, intra-cellular scale to the scale of the whole organism), and a large number of interacting physiological processes (see the so-called "hallmarks of cancer"), several questions are not fully understood. Among these, we want to focus on the most clinically relevant ones, such as the general laws governing tumor growth and the development of metastases (secondary tumors, responsible of 90% of the deaths from a solid cancer). In this context, it is thus challenging to exploit the diversity of the data available in experimental settings (such as *in vitro* tumor spheroids or *in vivo* mice experiments) in order to improve our understanding of the disease and its dynamics, which in turn lead to validation, refinement and better tuning of the macroscopic models used in the axes 1 and 2 for clinical applications.

In recent years, several new findings challenged the classical vision of the metastatic development biology, in particular by the discovery of organism-scale phenomena that are amenable to a dynamical description in terms of mathematical models based on differential equations. These include the angiogenesis-mediated distant inhibition of secondary tumors by a primary tumor the pre-metastatic niche or the self-seeding phenomenon. Building a general, cancer type specific, comprehensive theory that would integrate these dynamical processes

remains an open challenge. On the therapeutic side, recent studies demonstrated that some drugs (such as the Sunitinib), while having a positive effect on the primary tumor (reduction of the growth), could *accelerate* the growth of the metastases. Moreover, this effect was found to be scheduling-dependent. Designing better ways to use this drug in order to control these phenomena is another challenge. In the context of combination therapies, the question of the *sequence* of administration between the two drugs is also particularly relevant.

One of the technical challenge that we need to overcome when dealing with biological data is the presence of potentially very large inter-animal (or inter-individual) variability.

Starting from the available multi-modal data and relevant biological or therapeutic questions, our purpose is to develop adapted mathematical models (*i.e.* identifiable from the data) that recapitulate the existing knowledge and reduce it to its more fundamental components, with two main purposes:

1. to generate quantitative and empirically testable predictions that allow to assess biological hypotheses or
2. to investigate the therapeutic management of the disease and assist preclinical studies of anti-cancerous drug development.

We believe that the feedback loop between theoretical modeling and experimental studies can help to generate new knowledge and improve our predictive abilities for clinical diagnosis, prognosis, and therapeutic decision. Let us note that the first point is in direct link with the axes 1 and 2 of the team since it allows us to experimentally validate the models at the biological scale (*in vitro* and *in vivo* experiments) for further clinical applications.

More precisely, we first base ourselves on a thorough exploration of the biological literature of the biological phenomena we want to model: growth of tumor spheroids, *in vivo* tumor growth in mice, initiation and development of the metastases, effect of anti-cancerous drugs. Then we investigate, using basic statistical tools, the data we dispose, which can range from: spatial distribution of heterogeneous cell population within tumor spheroids, expression of cell makers (such as green fluorescent protein for cancer cells or specific antibodies for other cell types), bioluminescence, direct volume measurement or even intra-vital images obtained with specific imaging devices. According to the data type, we further build dedicated mathematical models that are based either on PDEs (when spatial data is available, or when time evolution of a structured density can be inferred from the data, for instance for a population of tumors) or ODEs (for scalar longitudinal data). These models are confronted to the data by two principal means:

1. when possible, experimental assays can give a direct measurement of some parameters (such as the proliferation rate or the migration speed) or
2. statistical tools to infer the parameters from observables of the model.

This last point is of particular relevance to tackle the problem of the large inter-animal variability and we use adapted statistical tools such as the mixed-effects modeling framework.

Once the models are shown able to describe the data and are properly calibrated, we use them to test or simulate biological hypotheses. Based on our simulations, we then aim at proposing to our biological collaborators new experiments to confirm or infirm newly generated hypotheses, or to test different administration protocols of the drugs. For instance, in a collaboration with the team of the professor Andreas Bikfalvi (Laboratoire de l'Angiogénèse et du Micro-environnement des Cancers, Inserm, Bordeaux), based on confrontation of a mathematical model to multi-modal biological data (total number of cells in the primary and distant sites and MRI), we could demonstrate that the classical view of metastatic dissemination and development (one metastasis is born from one cell) was probably inaccurate, in mice grafted with metastatic kidney tumors. We then proposed that metastatic germs could merge or attract circulating cells. Experiments involving cells tagged with two different colors are currently performed in order to confirm or infirm this hypothesis.

Eventually, we use the large amount of temporal data generated in preclinical experiments for the effect of anti-cancerous drugs in order to design and validate mathematical formalisms translating the biological mechanisms of action of these drugs for application to clinical cases, in direct connection with the axis 1. We have a special focus on targeted therapies (designed to specifically attack the cancer cells while sparing the

healthy tissue) such as the Sunitinib. This drug is indeed indicated as a first line treatment for metastatic renal cancer and we plan to conduct a translational study coupled between A. Bikfalvi's laboratory and medical doctors, F. Cornelis (radiologist) and A. Ravaud (head of the medical oncology department).

MYCENAE Project-Team

3. Research Program

3.1. Project team positioning

The main goal of MYCENAE is to address crucial questions arising from both Neuroendocrinology and Neuroscience from a mathematical perspective. The choice and subsequent study of appropriate mathematical formalisms to investigate these dynamics is at the core of MYCENAE's scientific foundations: slow-fast dynamical systems with multiple time scales, mean-field approaches subject to limit-size and stochastic effects, transport-like partial differential equations (PDE) and stochastic individual based models (SIBM).

The scientific positioning of MYCENAE is on the way between Mathematical Biology and Mathematics: we are involved both in the modeling of physiological processes and in the deep mathematical analysis of models, whether they be (i) models developed (or under development) within the team (ii) models developed by collaborating teams or (iii) benchmark models from the literature.

Our research program is grounded on previous results obtained in the framework of the **REGATE** (REGulation of the GonAdoTropE axis) Large Scale Initiative Action and the SISYPHE project team on the one hand, and the **Mathematical Neuroscience Team** in the **Center for Interdisciplinary Research in Biology** (Collège de France), on the other hand. Several of our research topics are related to the study and generalization of 2 master models: a 4D, multiscale in time, nonlinear model based on coupled FitzHugh-Nagumo dynamics that has proved to be a fruitful basis for the study of the complex oscillations in hypothalamic GnRH dynamics [27], [26], and a n D, multiscale in space, system of weakly-coupled non conservative transport equations that underlies our approach of gonadal cell dynamics [28],[8]. Most of our topics in mathematical neuroscience deal with the study of complex oscillatory behaviors exhibited either by single neurons or as emergent macroscopic properties of neural networks, from both a deterministic and stochastic viewpoint.

3.2. Numerical and theoretical studies of slow-fast systems with complex oscillations

In dynamical systems with at least three state variables, the presence of different time scales favors the appearance of complex oscillatory solutions. In this context, with (at least) two slow variables MixedMode Oscillations (MMO) dynamics can arise. MMOs are small and large amplitude oscillations combined in a single time series. The last decade has witnessed a significant amount of research on this topic, including studies of folded singularities, construction of MMOs using folded singularities in combination with global dynamics, effects of additional time scales, onset of MMOs via singular Hopf bifurcations, as well as generalization to higher dimensions. In the same period, many applications to neuroscience emerged [9]. On the other hand, bursting oscillations, another prototype of complex oscillations can occur in systems with (at least) two fast variables. Bursting has been observed in many biological contexts, in particular in the dynamics of pancreatic cells, neurons, and other excitable cells. In neuronal dynamics a burst corresponds to a series of spikes, interspersed with periods of quiescent behavior, called inter-burst intervals. We are interested in systems combining bursting, MMOs and canards. One of the interesting directions is torus canards, which are canard-like structures occurring in systems combining canard explosion with fast rotation [5]. Torus canards help understand transitions from spiking or MMO dynamics to bursting. Another study on the boundary of bursting and MMOs is the work of [36] on the so-called plateau bursting. A major challenge in this direction is to gain a complete understanding of the transition from “3 time scales” to “2 fast/ 1 slow” (bursting) and then to “1 fast/ 2 slow (MMOs)”. Also, a key challenge that we intend to tackle in the next few years is that of large dynamical systems with many fast and many slow variables, which additionally are changing in time and/or in phase space. We aim to pursue this research direction both at theoretical and computational level, using numerical continuation approaches based on the location of unstable trajectories by using fixed point methods, rather than simulation, to locate trajectories.

3.3. Non conservative transport equations for cell population dynamics

Models for physiologically-structured populations can be considered to derive from the so-called McKendrick-Von Foerster equation or renewal equation that has been applied and generalized in different applications of population dynamics, including ecology, epidemiology and cell biology. Renewal equations are PDE transport equations that are written so as to combine conservation laws (e.g. on the total number of individuals) with additional terms related to death or maturation, that blur the underlying overall balance law [33]. Renewal equations can be deployed only in contexts where a deterministic and continuous formalization is suitable to describe the populations. In the case of low or very low number of individuals, the probabilistic nature of the events driving the population dynamics has to be accounted for. In that context, the formalism initially developed in the framework of ecological modeling (see e.g. [30]) is particularly interesting since it can bridge the gap between branching processes and renewal equations for structured populations.

The development of ovarian follicles is a tightly-controlled physiological and morphogenetic process, that can be investigated from a middle-out approach starting at the cell level.

To describe the terminal stages of follicular development on a cell kinetics basis and account for the selection process operated amongst follicles, we have developed a multiscale model describing the cell density in each follicle, that can be roughly considered as a system of weakly-coupled, non conservative transport equations with controlled velocities and source term. Even if, in some sense, this model belongs to the class of renewal equations for structured populations, it owns a number of specificities that render its theoretical and numerical analysis particularly challenging: 2 structuring variables (per follicle, leading as a whole to $2nD$ system), control terms operating on the velocities and source term, and formulated from moments of the unknowns, discontinuities both in the velocities and density on internal boundaries of the domain representing the passage from one cell phase to another. On the theoretical ground, the well-posedness (existence and uniqueness of weak solutions with bounded initial data) has been established in [12], while associated control problems have been studied in the framework of hybrid optimal control [6]. On the numerical ground, the formalism dedicated to the simulation of these hyperbolic-like PDEs is that of finite volume method. Part of the numerical strategy consists in combining in the most efficient way low resolution numerical schemes (such as the first-order Godunov scheme), that tend to be diffusive, with high resolution schemes (such as the Lax Wendroff second-order scheme), that may engender oscillations in the vicinity of discontinuities [2], with a critical choice of the limiter functions. The 2D finite volume schemes are combined with adaptive mesh refinement through a multi-resolution method [4] and implemented in a problem-specific way on parallel architecture [1].

To describe the first stages of follicular development, which only involve a few cells, we call to a stochastic and discrete formalism. We have designed an individual-based, stochastic model [7] embedding specific laws of morphodynamics, which leads to a multiscale model where individual cells are endowed with a non-zero size and occupy space partitions of predefined sizes, which can bear a limit rate of overcrowding.

3.4. Macroscopic limits of stochastic neural networks and neural fields

The coordinated activity of the cortex is the result of the interactions between a very large number of cells. Each cell is well described by a dynamical system, that receives non constant input which is the superposition of an external stimulus, noise and interactions with other cells. Most models describing the emergent behavior arising from the interaction of neurons in large-scale networks have relied on continuum limits ever since the seminal work of Wilson and Cowan and Amari [37], [25]. Such models tend to represent the activity of the network through a macroscopic variable, the population-averaged firing rate.

In order to rationally describe neural fields and more generally large cortical assemblies, one should yet base their approach on what is known of the microscopic neuronal dynamics. At this scale, the equation of the activity is a set of stochastic differential equations in interaction. Obtaining the equations of evolution of the effective mean-field from microscopic dynamics is a very complex problem which belongs to statistical physics. As in the case of the kinetic theory of gases, macroscopic states are defined by the limit of certain quantities as the network size tends to infinity. When such a limit theorem is proved, one can be ensured that large networks are well approximated by the obtained macroscopic system. Qualitative distinctions between

the macroscopic limit and finite-sized networks (finite-size effects), occurs in such systems. We have been interested in the relevant mathematical approaches dealing with macroscopic limits of stochastic neuronal networks, that are expressed in the form of a complex integro-differential stochastic implicit equations of McKean-Vlasov type including a new mathematical object, the spatially chaotic Brownian motion [15].

The major question consists in establishing the fundamental laws of the collective behaviors cortical assemblies in a number of contexts motivated by neuroscience, such as communication delays between cells [14], [13] or spatially extended areas, which is the main topic of our current research. In that case additional difficulties arise, since the connection between different neurons, as well as delays in communications, depend on space in a correlated way, leading to the singular dependence of the solutions in space, which is not measurable.

NUMED Project-Team

3. Research Program

3.1. Design of complex models

3.1.1. Project team positioning

The originality of our work is the quantitative description of propagation phenomena accounting for several time and spatial scales. Here, propagation has to be understood in a broad sense. This includes propagation of invasive species, chemotactic waves of bacteria, evolution of age structured populations ... Our main objectives are the quantitative calculation of macroscopic quantities as the rate of propagation, and microscopic distributions at the edge and the back of the front. These are essential features of propagation which are intimately linked in the long time dynamics.

3.1.2. Recent results

- Mixed evolution - propagation models:

Vincent Calvez is studying propagation phenomena at the mesoscale, including travelling waves, accelerating fronts, evolutionary adaptation of a population to a changing environment. The common feature between these projects is the strong heterogeneity inside the propagating front. It is a great mathematical challenge to be able to keep track of this heterogeneity throughout the mathematical analysis.

This research is structured in several axes, which all belong to the ERC starting grant MESOPROBIO (2015-2020).

- Modeling and data analysis of species' invasion:

We are currently developing two directions of research on the case study of cane toads invasion in Northern Australia. The first direction is based on a dataset provided by Australian group of biologists. They recorded individual trajectories of individuals at a fixed location during ten consecutive years (2005-2015). We are using our expertise about waves of expansion in kinetic transport equations in order to calibrate some dedicated mesoscopic models on data. So far we experienced some difficulties because any approach severely underestimate the real speed of expansion. We are revisiting the problem based on this apparent counter-intuitive results. This was the purpose of Nils Caillerie's PhD thesis (defended July 2017).

The second direction is more theoretical. Namely we are seeking the true rate of expansion of a minimal model for wave front acceleration inspired from the cane toad study. This mathematical question appeared to much more difficult than expected. We are attacking it using a mixture of theoretical and numerical computations (in collaboration with T. Dumont, C. Henderson, S. Mirrahimi and O. Turanova).

- Analysis of subdiffusive processes:

H. Berry, V. Calvez and T. Lepoutre co-supervised the PhD of Alvaro Mateos Gonzalez. He studied large scale asymptotics of age-structured sub-diffusive models.

- Concentration waves of chemotactic bacteria at the mesoscale:

Vincent Calvez achieved a long standing goal in 2015, by proving the existence of traveling wave in a coupled kinetic/parabolic system of equations modeling bacteria chemotaxis. This project was grounded on biological experiments. There was several parallel sub-projects, including the PhD thesis of Emeric Bouin, and Nils Caillerie, as well as Hélène

Hivert's post-doctoral project. VC is also collaborating with Laurent Gosse and Monika Twarogowska on well balanced schemes for kinetic traveling waves.

This project led to new types of non-local Hamilton-Jacobi equations, including a collaboration with E. Bouin, G. Nadin and E. Grenier.

- Adaptation of a population to a changing environment:

Vincent Calvez moved recently to evolutionary biology, by some collaboration with biologists at Montpellier. The goal is to analyse quantitative genetics model using modern tools of PDE analysis, in particular WKB expansions, as initiated by Perthame and co-authors in 2005. We made two breakthroughs: firstly, we obtained quantitative results for age-structured population models, which add a level of complexity (the population is described by a phenotypic trait and age of individuals). Secondly, we realized that WKB expansions, which are well designed for linear equations (here, asexual mode of reproduction), could be extended to some non linear equations, including some sexual mode of reproduction. This paves the way for new mathematical challenges, as the asymptotic analysis requires new tools, and new quantitative results in evolutionary biology.

Vincent Calvez is currently at UBC, Vancouver, over the period 08/2017-03/2018, to consolidate the last project and initiate new collaborations there.

- Inviscid limit of Navier Stokes equations.

The question of the behavior of solutions of Navier Stokes equations in a bounded domain as the viscosity goes to 0 is a classical and highly difficult open question in Fluid Mechanics. A small boundary layer, called Prandtl layer, appears near the boundary, which turns out to be unstable if the viscosity is small enough. The stability analysis of this boundary layer is highly technical and remained open since the first formal analysis in the 1940's by physicists like Orr, Sommerfeld, Tollmien, Schlichting or Lin. E. Grenier recently made a complete mathematical analysis of this spectral problem, in collaboration with T. Nguyen and Y. Guo. We rigorously proved that any shear layer is spectrally and linearly unstable if the viscosity is small enough, which is the first mathematical result in that field. We also get some preliminary nonlinear results. A book on this subject is in preparation, already accepted by Springer.

- Qualitative properties.

Sometimes a beautiful mathematical question arises from modeling. For instance the question to know whether the apparent size of a tumor on a MRI can decrease long after a radiotherapy is linked to the following mathematical question: if u is a solution of the classical KPP equation, does it satisfies $\partial_t u > 0$ everywhere provided t is large enough ? Despite its simplicity, this natural question appeared open and delicate. With the help of F. Hamel, E. Grenier managed to prove that this is true (to be published in JMPA) [13].

- Numerical analysis of complex fluids: the example of avalanches.

This deals with the development of numerical schemes for viscoplastic materials (namely with Bingham or Herschell-Bulkley laws). Recently, with other colleagues, Paul Vigneaux finished the design of the first 2D well-balanced finite volume scheme for a shallow viscoplastic model. It is illustrated on the famous Tacconnaz avalanche path in the Mont-Blanc, Chamonix, in the case of dense snow avalanches. The scheme deals with general Digital Elevation Model (DEM) topographies, wet/dry fronts and is designed to compute precisely the stopping state of avalanches, a crucial point of viscoplastic flows which are able to rigidify [21].

3.1.3. Collaborations

- Mixed evolution: N. Bournaveas (Edinburgh), B. Perthame (Paris 6), C. Schmeiser (Vienna), P. Silberzan (Institut Curie), S. Mirrahimi (Toulouse).
- Inviscid limit of Navier Stokes equations: Brown University (Y. Guo, B. Pausader), Penn State University (T. Nguyen), Orsay University (F. Rousset).

- Qualitative properties: F. Hamel (Marseille University).
- Numerical analysis of complex fluids: Enrique D. Fernandez - Nieto (Univ. de Sevilla, Spain), Jose Maria Gallardo (Univ. de Malaga, Spain).

3.2. Parametrization of complex systems

3.2.1. Project-team positioning

Clinical data are often sparse: we have few data per patient. The number of data is of the order of the number of parameters. In this context, a natural way to parametrize complex models with real world clinical data is to use a Bayesian approach, namely to try to find the distribution of the model parameters in the population, rather than to try to identify the parameters of every single patient. This approach has been pioneered in the 90's by the Nonmem software, and has been much improved thanks to Marc Lavielle in the 2000's. Refined statistical methods, called SAEM, have been tuned and implemented in commercial softwares like Monolix.

3.2.2. Recent results

The main problem when we try to parametrize clinical data using complex systems is the computational time. One single evaluation of the model can be costly, in particular if this model involves partial differential equations, and SAEM algorithm requires hundreds of thousands of single evaluations. The time cost is then too large, in particular because SAEM may not be parallelized.

To speed up the evaluation of the complex model, we replace it by an approximate one, or so called metamodel, constructed by interpolation of a small number of its values. We therefore combine the classical SAEM algorithm with an interpolation step, leading to a strong acceleration. Interpolation can be done through a precomputation step on a fixed grid, or through a more efficient kriging step. The interpolation grid or the kriging step may be improved during SAEM algorithm in an iterative way in order to get accurate evaluations of the complex system only in the domain of interest, namely near the clinical values [14],[15].

We applied these new algorithms to synthetic data and are currently using them on glioma data. We are also currently trying to prove the convergence of the corresponding algorithms. We will develop glioma applications in the next section.

Moreover E. Ollier in his PhD developed new strategies to distinguish various populations within a SAEM algorithm [23].

We have two long standing collaborations with Sanofi and Servier on parametrization issues:

- Servier: during a four years contract, we modelled the pkpd of new drugs and also study the combination and optimization of chimiotherapies.
- Sanofi: during a eight years contract, Emmanuel Grenier wrote a complete software devoted to the study of the degradation of vaccine. This software is used worldwide by Sanofi R&D teams in order to investigate the degradation of existing or new vaccines and to study their behavior when they are heated. This software has been used on flu, dengue and various other diseases.

3.2.3. Collaborations

- Academic collaborations: A. Leclerc Samson (Grenoble University)
- Medical collaborations: Dr Ducray (Centre Léon Bérard, Lyon) and Dr Sujobert (Lyon Sud Hospital)
- Industrial contracts: we used parametrization and treatment improvement techniques for Servier (four years contract, on cancer drug modeling and optimization) and Sanofi (long standing collaboration)

3.3. Multiscale models in oncology

3.3.1. Project-team positioning

Cancer modeling is the major topic of several teams in France and Europe, including Mamba, Monc and Asclepios to quote only a few Inria teams. These teams try to model metastasis, tumoral growth, vascularisation through angiogenesis, or to improve medical images quality. Their approaches are based on dynamical systems, partial differential equations, or on special imagery techniques.

Numed focuses on the link between very simple partial differential equations models, like reaction diffusion models, and clinical data.

3.3.2. Results

We managed to build a clinical database, which gathers clinical data from about twenty different patients which were treated in Lyon Hospital. For each patient, we have five to twenty different MRIs. Each MRI has been segmented manually by A. Peters, and checked by Dr Ducray. There are about two hundred different MRIs.

With this database, we can try to parametrize various models of glioma, thanks to the population parametrization techniques developed under axis 2. We are currently trying to parametrize a simple reaction diffusion model for the decay of glioma after chemotherapy [7].

3.4. Stroke

3.4.1. Project-team positioning

Many teams work on medical imagery, in order to improve diffusion images or MRIs, or to improve their mathematical analysis. But there are few academic teams working on ischemic stroke modeling in Europe, mainly because there are few available clinical data for a given patient. Except routine imagery, often with low definition because it is taken very quickly, in an emergency context, there are no biochemical data. The follow up is also very sparse.

However the domain is very rich. Ischemic stroke involves blood flow, ionic exchanges, cell swelling, cell death (including necrosis and apoptosis), reperfusion, free radicals ... It is possible to build very detailed descriptive models of stroke, however it is not possible to parametrize these models using clinical data. Only a crude parametrization using bibliographical data or ad hoc parameters is possible. We followed this path at the beginning of Numed. Now we are focused on clinical data, namely diffusion and perfusion maps.

The scientific challenge is to try to predict the outcome of the patient, starting from the initial clinical images, obtained when the patient enters the hospital. By *outcome*, we mean the final size of the dead area, or patient abilities (speech, walk, standing). It is particularly important to try to know whether clinicians have to give some particular drug to try to reopen blood flow, or whether this reopening would be harmful for the patient.

3.4.2. Results

Through the PhD thesis of Mathilde Giacalone, we built a strong collaboration with an imagery team of the Creatis lab (Lyon I). This team is very strong in the analysis of clinical images of stroke and in particular on perfusion. It closely works with clinicians (Pr Nighoghossian at Lyon Bron hospitals). We have therefore built a strong connection with clinicians and image specialists.

Thanks to the work of Mathilde Giacalone, we have now at hand a large database of clinical data (more than 50 patients). For each of these patients we have at hand their perfusion and diffusion maps at their entry, together with diffusion one week and one month later. All these images have been checked and segmented. All the data have been supervised by a clinician. The database is therefore very reliable. Its setup took almost two years. This database will grow with time, according to the patients treated in Lyon.

Mathilde Giacalone worked on new algorithms to improve perfusion images. A perfusion image describes the blood flow in the brain. We also fulfilled a preliminary analysis to know what kind of image (there are several protocols to get perfusion images) is the best to predict the outcome of the patient, using information theory.

REO Project-Team

3. Research Program

3.1. Multiphysics modeling

In large vessels and in large bronchi, blood and air flows are generally supposed to be governed by the incompressible Navier-Stokes equations. Indeed in large arteries, blood can be supposed to be Newtonian, and at rest air can be modeled as an incompressible fluid. The cornerstone of the simulations is therefore a Navier-Stokes solver. But other physical features have also to be taken into account in simulations of biological flows, in particular fluid-structure interaction in large vessels and transport of sprays, particles or chemical species.

3.1.1. Fluid-structure interaction

Fluid-structure coupling occurs both in the respiratory and in the circulatory systems. We focus mainly on blood flows since our work is more advanced in this field. But the methods developed for blood flows could be also applied to the respiratory system.

Here “fluid-structure interaction” means a coupling between the 3D Navier-Stokes equations and a 3D (possibly thin) structure in large displacements.

The numerical simulations of the interaction between the artery wall and the blood flows raise many issues: (1) the displacement of the wall cannot be supposed to be infinitesimal, geometrical nonlinearities are therefore present in the structure and the fluid problem have to be solved on a moving domain (2) the densities of the artery walls and the blood being close, the coupling is strong and has to be tackled very carefully to avoid numerical instabilities, (3) “naïve” boundary conditions on the artificial boundaries induce spurious reflection phenomena.

Simulation of valves, either at the outflow of the cardiac chambers or in veins, is another example of difficult fluid-structure problems arising in blood flows. In addition, very large displacements and changes of topology (contact problems) have to be handled in those cases.

Due to stability reasons, it seems impossible to successfully apply in hemodynamics the explicit coupling schemes used in other fluid-structure problems, like aeroelasticity. As a result, fluid-structure interaction in biological flows raise new challenging issues in scientific computing and numerical analysis : new schemes have to be developed and analyzed.

We have proposed and analyzed over the last few years several efficient fluid-structure interaction algorithms. This topic remains very active. We are now using these algorithms to address inverse problems in blood flows to make patient specific simulations (for example, estimation of artery wall stiffness from medical imaging).

3.1.2. Aerosol

Complex two-phase fluids can be modeled in many different ways. Eulerian models describe both phases by physical quantities such as the density, velocity or energy of each phase. In the mixed fluid-kinetic models, the biphasic fluid has one dispersed phase, which is constituted by a spray of droplets, with a possibly variable size, and a continuous classical fluid.

This type of model was first introduced by Williams [57] in the frame of combustion. It was later used to develop the Kiva code [45] at the Los Alamos National Laboratory, or the Hesione code [52], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [49], therapeutic sprays, *etc.* One of the interests of such a model is that various phenomena on the droplets can be taken into account with an accurate precision: collision, breakups, coagulation, vaporization, chemical reactions, *etc.*, at the level of the droplets.

The model usually consists in coupling a kinetic equation, that describes the spray through a probability density function, and classical fluid equations (typically Navier-Stokes). The numerical solution of this system relies on the coupling of a method for the fluid equations (for instance, a finite volume method) with a method fitted to the spray (particle method, Monte Carlo).

We are mainly interested in modeling therapeutic sprays either for local or general treatments. The study of the underlying kinetic equations should lead us to a global model of the ambient fluid and the droplets, with some mathematical significance. Well-chosen numerical methods can give some tracks on the solutions behavior and help to fit the physical parameters which appear in the models.

3.2. Multiscale modeling

Multiscale modeling is a necessary step for blood and respiratory flows. In this section, we focus on blood flows. Nevertheless, similar investigations are currently carried out on respiratory flows.

3.2.1. Arterial tree modeling

Problems arising in the numerical modeling of the human cardiovascular system often require an accurate description of the flow in a specific sensible subregion (carotid bifurcation, stented artery, *etc.*). The description of such local phenomena is better addressed by means of three-dimensional (3D) simulations, based on the numerical approximation of the incompressible Navier-Stokes equations, possibly accounting for compliant (moving) boundaries. These simulations require the specification of boundary data on artificial boundaries that have to be introduced to delimit the vascular district under study. The definition of such boundary conditions is critical and, in fact, influenced by the global systemic dynamics. Whenever the boundary data is not available from accurate measurements, a proper boundary condition requires a mathematical description of the action of the reminder of the circulatory system on the local district. From the computational point of view, it is not affordable to describe the whole circulatory system keeping the same level of detail. Therefore, this mathematical description relies on simpler models, leading to the concept of *geometrical multiscale* modeling of the circulation [53]. The underlying idea consists in coupling different models (3D, 1D or 0D) with a decreasing level of accuracy, which is compensated by their decreasing level of computational complexity.

The research on this topic aims at providing a correct methodology and a mathematical and numerical framework for the simulation of blood flow in the whole cardiovascular system by means of a geometric multiscale approach. In particular, one of the main issues will be the definition of stable coupling strategies between 3D and reduced order models.

To model the arterial tree, a standard way consists of imposing a pressure or a flow rate at the inlet of the aorta, *i.e.* at the network entry. This strategy does not allow to describe important features as the overload in the heart caused by backward traveling waves. Indeed imposing a boundary condition at the beginning of the aorta artificially disturbs physiological pressure waves going from the arterial tree to the heart. The only way to catch this physiological behavior is to couple the arteries with a model of heart, or at least a model of left ventricle.

A constitutive law for the myocardium, controlled by an electrical command, has been developed in the CardioSense3D project⁰. One of our objectives is to couple artery models with this heart model.

A long term goal is to achieve 3D simulations of a system including heart and arteries. One of the difficulties of this very challenging task is to model the cardiac valves. To this purpose, we investigate a mix of arbitrary Lagrangian Eulerian and fictitious domain approaches or x-fem strategies, or simplified valve models based on an immersed surface strategy.

⁰<http://www-sop.inria.fr/CardioSense3D/>

3.2.2. Heart perfusion modeling

The heart is the organ that regulates, through its periodical contraction, the distribution of oxygenated blood in human vessels in order to nourish the different parts of the body. The heart needs its own supply of blood to work. The coronary arteries are the vessels that accomplish this task. The phenomenon by which blood reaches myocardial heart tissue starting from the blood vessels is called in medicine perfusion. The analysis of heart perfusion is an interesting and challenging problem. Our aim is to perform a three-dimensional dynamical numerical simulation of perfusion in the beating heart, in order to better understand the phenomena linked to perfusion. In particular the role of the ventricle contraction on the perfusion of the heart is investigated as well as the influence of blood on the solid mechanics of the ventricle. Heart perfusion in fact implies the interaction between heart muscle and blood vessels, in a sponge-like material that contracts at every heartbeat via the myocardium fibers.

Despite recent advances on the anatomical description and measurements of the coronary tree and on the corresponding physiological, physical and numerical modeling aspects, the complete modeling and simulation of blood flows inside the large and the many small vessels feeding the heart is still out of reach. Therefore, in order to model blood perfusion in the cardiac tissue, we must limit the description of the detailed flows at a given space scale, and simplify the modeling of the smaller scale flows by aggregating these phenomena into macroscopic quantities, by some kind of “homogenization” procedure. To that purpose, the modeling of the fluid-solid coupling within the framework of porous media appears appropriate.

Poromechanics is a simplified mixture theory where a complex fluid-structure interaction problem is replaced by a superposition of both components, each of them representing a fraction of the complete material at every point. It originally emerged in soils mechanics with the work of Terzaghi [56], and Biot [46] later gave a description of the mechanical behavior of a porous medium using an elastic formulation for the solid matrix, and Darcy’s law for the fluid flow through the matrix. Finite strain poroelastic models have been proposed (see references in [47]), albeit with *ad hoc* formulations for which compatibility with thermodynamics laws and incompressibility conditions is not established.

3.2.3. Tumor and vascularization

The same way the myocardium needs to be perfused for the heart to beat, when it has reached a certain size, tumor tissue needs to be perfused by enough blood to grow. It thus triggers the creation of new blood vessels (angiogenesis) to continue to grow. The interaction of tumor and its micro-environment is an active field of research. One of the challenges is that phenomena (tumor cell proliferation and death, blood vessel adaptation, nutrient transport and diffusion, etc) occur at different scales. A multi-scale approach is thus being developed to tackle this issue. The long term objective is to predict the efficiency of drugs and optimize therapy of cancer.

3.2.4. Respiratory tract modeling

We aim at developing a multiscale model of the respiratory tract. Intraparenchymal airways distal from generation 7 of the tracheobronchial tree (TBT), which cannot be visualized by common medical imaging techniques, are modeled either by a single simple model or by a model set according to their order in TBT. The single model is based on straight pipe fully developed flow (Poiseuille flow in steady regimes) with given alveolar pressure at the end of each compartment. It will provide boundary conditions at the bronchial ends of 3D TBT reconstructed from imaging data. The model set includes three serial models. The generation down to the pulmonary lobule will be modeled by reduced basis elements. The lobular airways will be represented by a fractal homogenization approach. The alveoli, which are the gas exchange loci between blood and inhaled air, inflating during inspiration and deflating during expiration, will be described by multiphysics homogenization.

SISTM Project-Team

3. Research Program

3.1. Mechanistic modelling

When studying the dynamics of a given marker, say the HIV concentration in the blood (HIV viral load), one can for instance use descriptive models summarising the dynamics over time in term of slopes of the trajectories [57]. These slopes can be compared between treatment groups or according to patients' characteristics. Another way for analysing these data is to define a mathematical model based on the biological knowledge of what drives HIV dynamics. In this case, it is mainly the availability of target cells (the CD4+ T lymphocytes), the production and death rates of infected cells and the clearance of the viral particles that impact the dynamics. Then, a mathematical model most often based on ordinary differential equations (ODE) can be written [50]. Estimating the parameters of this model to fit observed HIV viral load gave a crucial insight in HIV pathogenesis as it revealed the very short half-life of the virions and infected cells and therefore a very high turnover of the virus, making mutations a very frequent event [49].

Having a good mechanistic model in a biomedical context such as HIV infection opens doors to various applications beyond a good understanding of the data. Global and individual predictions can be excellent because of the external validity of a model based on main biological mechanisms. Control theory may serve for defining optimal interventions or optimal designs to evaluate new interventions [42]. Finally, these models can capture explicitly the complex relationship between several processes that change over time and may therefore challenge other proposed approaches such as marginal structural models to deal with causal associations in epidemiology [41].

Therefore, we postulate that this type of model could be very useful in the context of our research that is in complex biological systems. The definition of the model needs to identify the parameter values that fit the data. In clinical research this is challenging because data are sparse, and often unbalanced, coming from populations of subjects. A substantial inter-individual variability is always present and needs to be accounted as this is the main source of information. Although many approaches have been developed to estimate the parameters of non-linear mixed models [53], [60], [45], [51], [46], [59], the difficulty associated with the complexity of ODE models and the sparsity of the data leading to identifiability issues need further research.

3.2. High dimensional data

With the availability of omics data such as genomics (DNA), transcriptomics (RNA) or proteomics (proteins), but also other types of data, such as those arising from the combination of large observational databases (e.g. in pharmacoepidemiology or environmental epidemiology), high-dimensional data have become increasingly common. Use of molecular biological technics such as Polymerase Chain Reaction (PCR) allows for amplification of DNA or RNA sequences. Nowadays, microarray and Next Generation Sequencing (NGS) techniques give the possibility to explore very large portions of the genome. Furthermore, other assays have also evolved, and traditional measures such as cytometry or imaging have become new sources of big data. Therefore, in the context of HIV research, the dimension of the datasets has much grown in term of number of variables per individual than in term of number of included patients although this latter is also growing thanks to the multi-cohort collaborations such as CASCADE or COHERE organized in the EuroCoord network⁰. As an example, in a recent phase 1/2 clinical trial evaluating the safety and the immunological response to a dendritic cell-based HIV vaccine, 19 infected patients were included. Bringing together data on cell count, cytokine production, gene expression and viral genome change led to a 20 Go database [56]. This is far from big databases faced in other areas but constitutes a revolution in clinical research where clinical trials of hundred of patients sized few hundred of Ko at most. Therefore, more than the storage and calculation capacities, the challenge is the comprehensive analysis of these datasets.

⁰see online at <http://www.eurocoord.net>

The objective is either to select the relevant information or to summarize it for understanding or prediction purposes. When dealing with high dimensional data, the methodological challenge arises from the fact that datasets typically contain many variables, much more than observations. Hence, multiple testing is an obvious issue that needs to be taken into account [54]. Furthermore, conventional methods, such as linear models, are inefficient and most of the time even inapplicable. Specific methods have been developed, often derived from the machine learning field, such as regularization methods [58]. The integrative analysis of large datasets is challenging. For instance, one may want to look at the correlation between two large scale matrices composed by the transcriptome in the one hand and the proteome on the other hand [47]. The comprehensive analysis of these large datasets concerning several levels from molecular pathways to clinical response of a population of patients needs specific approaches and a very close collaboration with the providers of data that is the immunologists, the virologists, the clinicians...

XPOP Project-Team

3. Research Program

3.1. Scientific positioning

"Interfaces" is the defining characteristic of XPOP:

The interface between statistics, probability and numerical methods. Mathematical modelling of complex biological phenomena require to combine numerical, stochastic and statistical approaches. The CMAP is therefore the right place to be for positioning the team at the interface between several mathematical disciplines.

The interface between mathematics and the life sciences. The goal of XPOP is to bring the right answers to the right questions. These answers are mathematical tools (statistics, numerical methods, etc.), whereas the questions come from the life sciences (pharmacology, medicine, biology, etc.). This is why the point of XPOP is not to take part in mathematical projects only, but also pluridisciplinary ones.

The interface between mathematics and software development. The development of new methods is the main activity of XPOP. However, new methods are only useful if they end up being implemented in a software tool. A strong partnership with Lixoft (the spin-off company who continue developing MONOLIX) is indispensable to maintaining this positioning.

3.2. The mixed-effects models

Mixed-effects models are statistical models with both fixed effects and random effects. They are well-adapted to situations where repeated measurements are made on the same individual/statistical unit.

Consider first a single subject i of the population. Let $y_i = (y_{ij}, 1 \leq j \leq n_i)$ be the vector of observations for this subject. The model that describes the observations y_i is assumed to be a parametric probabilistic model: let $p_Y(y_i; \psi_i)$ be the probability distribution of y_i , where ψ_i is a vector of parameters.

In a population framework, the vector of parameters ψ_i is assumed to be drawn from a population distribution $p_\Psi(\psi_i; \theta)$ where θ is a vector of population parameters.

Then, the probabilistic model is the joint probability distribution

$$p(y_i, \psi_i; \theta) = p_Y(y_i | \psi_i) p_\Psi(\psi_i; \theta) \quad (19)$$

To define a model thus consists in defining precisely these two terms.

In most applications, the observed data y_i are continuous longitudinal data. We then assume the following representation for y_i :

$$y_{ij} = f(t_{ij}, \psi_i) + g(t_{ij}, \psi_i) \varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i. \quad (20)$$

Here, y_{ij} is the observation obtained from subject i at time t_{ij} . The residual errors (ε_{ij}) are assumed to be standardized random variables (mean zero and variance 1). The residual error model is represented by function g in model (2).

Function f is usually the solution to a system of ordinary differential equations (pharmacokinetic/pharmacodynamic models, etc.) or a system of partial differential equations (tumor growth, respiratory system, etc.). This component is a fundamental component of the model since it defines the prediction of the observed kinetics for a given set of parameters.

The vector of individual parameters ψ_i is usually function of a vector of population parameters ψ_{pop} , a vector of random effects $\eta_i \sim \mathcal{N}(0, \Omega)$, a vector of individual covariates c_i (weight, age, gender, ...) and some fixed effects β .

The joint model of y and ψ depends then on a vector of parameters $\theta = (\psi_{\text{pop}}, \beta, \Omega)$.

3.3. Computational Statistical Methods

Central to modern statistics is the use of probabilistic models. To relate these models to data requires the ability to calculate the probability of the observed data: the likelihood function, which is central to most statistical methods and provides a principled framework to handle uncertainty.

The emergence of computational statistics as a collection of powerful and general methodologies for carrying out likelihood-based inference made complex models with non-standard data accessible to likelihood, including hierarchical models, models with intricate latent structure, and missing data.

In particular, algorithms previously developed by POPIX for mixed effects models, and today implemented in several software tools (especially MONOLIX) are part of these methods:

- the adaptive Metropolis-Hastings algorithm allows one to sample from the conditional distribution of the individual parameters $p(\psi_i | y_i; c_i, \theta)$,
- the SAEM algorithm is used to maximize the observed likelihood $\mathcal{L}(\theta; y) = p(y; \theta)$,
- Importance Sampling Monte Carlo simulations provide an accurate estimation of the observed log-likelihood $\log(\mathcal{L}(\theta; y))$.

Computational statistics is an area which remains extremely active today. Recently, one can notice that the incentive for further improvements and innovation comes mainly from three broad directions: the high dimensional challenge, the quest for adaptive procedures that can eliminate the cumbersome process of tuning "by hand" the settings of the algorithms and the need for flexible theoretical support, arguably required by all recent developments as well as many of the traditional MCMC algorithms that are widely used in practice.

Working in these three directions is a clear objective for XPOP.

3.4. Markov Chain Monte Carlo algorithms

While these Monte Carlo algorithms have turned into standard tools over the past decade, they still face difficulties in handling less regular problems such as those involved in deriving inference for high-dimensional models. One of the main problems encountered when using MCMC in this challenging settings is that it is difficult to design a Markov chain that efficiently samples the state space of interest.

The Metropolis-adjusted Langevin algorithm (MALA) is a Markov chain Monte Carlo (MCMC) method for obtaining random samples from a probability distribution for which direct sampling is difficult. As the name suggests, MALA uses a combination of two mechanisms to generate the states of a random walk that has the target probability distribution as an invariant measure:

1. new states are proposed using Langevin dynamics, which use evaluations of the gradient of the target probability density function;
2. these proposals are accepted or rejected using the Metropolis-Hastings algorithm, which uses evaluations of the target probability density (but not its gradient).

Informally, the Langevin dynamics drives the random walk towards regions of high probability in the manner of a gradient flow, while the Metropolis-Hastings accept/reject mechanism improves the mixing and convergence properties of this random walk.

Several extensions of MALA have been proposed recently by several authors, including fMALA (fast MALA), AMALA (anisotropic MALA), MMALA (manifold MALA), position-dependent MALA (PMALA), ...

MALA and these extensions have demonstrated to represent very efficient alternative for sampling from high dimensional distributions. We therefore need to adapt these methods to general mixed effects models.

3.5. Parameter estimation

The Stochastic Approximation Expectation Maximization (SAEM) algorithm has shown to be extremely efficient for maximum likelihood estimation in incomplete data models, and particularly in mixed effects models for estimating the population parameters. However, there are several practical situations for which extensions of SAEM are still needed:

High dimensional model: a complex physiological model may have a large number of parameters (in the order of 100). Then several problems arise:

- when most of these parameters are associated with random effects, the MCMC algorithm should be able to sample, for each of the N individuals, parameters from a high dimensional distribution. Efficient MCMC methods for high dimensions are then required.
- Practical identifiability of the model is not ensured with a limited amount of data. In other words, we cannot expect to be able to properly estimate all the parameters of the model, including the fixed effects and the variance-covariance matrix of the random effects. Then, some random effects should be removed, assuming that some parameters do not vary in the population. It may also be necessary to fix the value of some parameters (using values from the literature for instance). The strategy to decide which parameters should be fixed and which random effects should be removed remains totally empirical. XPOP aims to develop a procedure that will help the modeller to take such decisions.

Large number of covariates: the covariate model aims to explain part of the inter-patient variability of some parameters. Classical methods for covariate model building are based on comparisons with respect to some criteria, usually derived from the likelihood (AIC, BIC), or some statistical test (Wald test, LRT, etc.). In other words, the modelling procedure requires two steps: first, all possible models are fitted using some estimation procedure (e.g. the SAEM algorithm) and the likelihood of each model is computed using a numerical integration procedure (e.g. Monte Carlo Importance Sampling); then, a model selection procedure chooses the "best" covariate model. Such a strategy is only possible with a reduced number of covariates, i.e., with a "small" number of models to fit and compare.

As an alternative, we are thinking about a Bayesian approach which consists of estimating simultaneously the covariate model and the parameters of the model in a single run. An (informative or uninformative) prior is defined for each model by defining a prior probability for each covariate to be included in the model. In other words, we extend the probabilistic model by introducing binary variables that indicate the presence or absence of each covariate in the model. Then, the model selection procedure consists of estimating and maximizing the conditional distribution of this sequence of binary variables. Furthermore, a probability can be associated to any of the possible covariate models.

This conditional distribution can be estimated using an MCMC procedure combined with the SAEM algorithm for estimating the population parameters of the model. In practice, such an approach can only deal with a limited number of covariates since the dimension of the probability space to explore increases exponentially with the number of covariates. Consequently, we would like to have methods able to find a small number of variables (from a large starting set) that influence certain parameters in populations of individuals. That means that, instead of estimating the conditional distribution of all the covariate models as described above, the algorithm should focus on the most likely ones.

Fixed parameters: it is quite frequent that some individual parameters of the model have no random component and are purely fixed effects. Then, the model may not belong to the exponential family anymore and the original version of SAEM cannot be used as it is. Several extensions exist:

- introduce random effects with decreasing variances for these parameters,
- introduce a prior distribution for these fixed effects,
- apply the stochastic approximation directly on the sequence of estimated parameters, instead of the sufficient statistics of the model.

None of these methods always work correctly. Furthermore, what are the pros and cons of these methods is not clear at all. Then, developing a robust methodology for such model is necessary.

Convergence toward the global maximum of the likelihood: convergence of SAEM can strongly depend on this initial guess when the observed likelihood has several local maxima. A kind of simulated annealing version of SAEM was previously developed and implemented in MONOLIX. The method works quite well in most situations but there is no theoretical justification and choosing the settings of this algorithm (i.e. how the temperature decreases during the iterations) remains empirical. A precise analysis of the algorithm could be very useful to better understand why it "works" in practice and how to optimize it.

Convergence diagnostic: Convergence of SAEM was theoretically demonstrated under very general hypothesis. Such result is important but of little interest in practice at the time to use SAEM in a finite amount of time, i.e. in a finite number of iterations. Some qualitative and quantitative criteria should be defined in order to both optimize the settings of the algorithm, detect a poor convergence of SAEM and evaluate the quality of the results in order to avoid using them unwisely.

3.6. Model building

Defining an optimal strategy for model building is far from easy because a model is the assembled product of numerous components that need to be evaluated and perhaps improved: the structural model, residual error model, covariate model, covariance model, etc.

How to proceed so as to obtain the best possible combination of these components? There is no magic recipe but an effort will be made to provide some qualitative and quantitative criteria in order to help the modeller for building his model.

The strategy to take will mainly depend on the time we can dedicate to building the model and the time required for running it. For relatively simple models for which parameter estimation is fast, it is possible to fit many models and compare them. This can also be done if we have powerful computing facilities available (e.g., a cluster) allowing large numbers of simultaneous runs.

However, if we are working on a standard laptop or desktop computer, model building is a sequential process in which a new model is tested at each step. If the model is complex and requires significant computation time (e.g., when involving systems of ODEs), we are constrained to limit the number of models we can test in a reasonable time period. In this context, it also becomes important to carefully choose the tasks to run at each step.

3.7. Model evaluation

Diagnostic tools are recognized as an essential method for model assessment in the process of model building. Indeed, the modeler needs to confront "his" model with the experimental data before concluding that this model is able to reproduce the data and before using it for any purpose, such as prediction or simulation for instance.

The objective of a diagnostic tool is twofold: first we want to check if the assumptions made on the model are valid or not ; then, if some assumptions are rejected, we want to get some guidance on how to improve the model.

As is the usual case in statistics, it is not because this "final" model has not been rejected that it is necessarily the "true" one. All that we can say is that the experimental data does not allow us to reject it. It is merely one of perhaps many models that cannot be rejected.

Model diagnostic tools are for the most part graphical, i.e., visual; we "see" when something is not right between a chosen model and the data it is hypothesized to describe. These diagnostic plots are usually based on the empirical Bayes estimates (EBEs) of the individual parameters and EBEs of the random effects: scatterplots of individual parameters versus covariates to detect some possible relationship, scatterplots of pairs of random effects to detect some possible correlation between random effects, plot of the empirical distribution of the random effects (boxplot, histogram,...) to check if they are normally distributed, ...

The use of EBEs for diagnostic plots and statistical tests is efficient with rich data, i.e. when a significant amount of information is available in the data for recovering accurately all the individual parameters. On the contrary, tests and plots can be misleading when the estimates of the individual parameters are greatly shrunk.

We propose to develop new approaches for diagnosing mixed effects models in a general context and derive formal and unbiased statistical tests for testing separately each feature of the model.

3.8. Missing data

The ability to easily collect and gather a large amount of data from different sources can be seen as an opportunity to better understand many processes. It has already led to breakthroughs in several application areas. However, due to the wide heterogeneity of measurements and objectives, these large databases often exhibit an extraordinary high number of missing values. Hence, in addition to scientific questions, such data also present some important methodological and technical challenges for data analyst.

Missing values occur for a variety of reasons: machines that fail, survey participants who do not answer certain questions, destroyed or lost data, dead animals, damaged plants, etc. Missing values are problematic since most statistical methods can not be applied directly on a incomplete data. Many progress have been made to properly handle missing values. However, there are still many challenges that need to be addressed in the future, that are crucial for the users.

- State of arts methods often consider the case of continuous or categorical data whereas real data are very often mixed. The idea is to develop a multiple imputation method based on a specific principal component analysis (PCA) for mixed data. Indeed, PCA has been used with success to predict (impute) the missing values. A very appealing property is the ability of the method to handle very large matrices with large amount of missing entries.
- The asymptotic regime underlying modern data is not any more to consider that the sample size increases but that both number of observations and number of variables are very large. In practice first experiments showed that the coverage properties of confidence areas based on the classical methods to estimate variance with missing values varied widely. The asymptotic method and the bootstrap do well in low-noise setting, but can fail when the noise level gets high or when the number of variables is much greater than the number of rows. On the other hand, the jackknife has good coverage properties for large noisy examples but requires a minimum number of variables to be stable enough.
- Inference with missing values is usually performed under the assumption of "Missing at Random" (MAR) values which means that the probability that a value is missing may depend on the observed data but does not depend on the missing value itself. In real data and in particular in data coming from clinical studies, both "Missing Non at Random" (MNAR) and MAR values occur. Taking into account in a proper way both types of missing values is extremely challenging but is worth investigating since the applications are extremely broad.

It is important to stress that missing data models are part of the general incomplete data models addressed by XPOP. Indeed, models with latent variables (i.e. non observed variables such as random effects in a mixed effects model), models with censored data (e.g. data below some limit of quantification) or models with dropout mechanism (e.g. when a subject in a clinical trial fails to continue in the study) can be seen as missing data models.