RESEARCH CENTER

FIELD
Digital Health, Biology and Earth

Activity Report 2015

Section Scientific Foundations

Edition: 2016-03-21

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 proteins [0], 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 change [0], 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.

[0] For 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.

[0] The 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 [3]. 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 relaxed [0]. 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 [6]. 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.

AMIB Project-Team

3. Research Program

3.1. RNA

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. Dynamic programming and complexity

Participants: Yann Ponty, Antoine Soulé.

*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 [56], the correction of pyrosequencing errors [48], and the efficient detection of multi-stable RNAs (riboswitches) [50], [51].

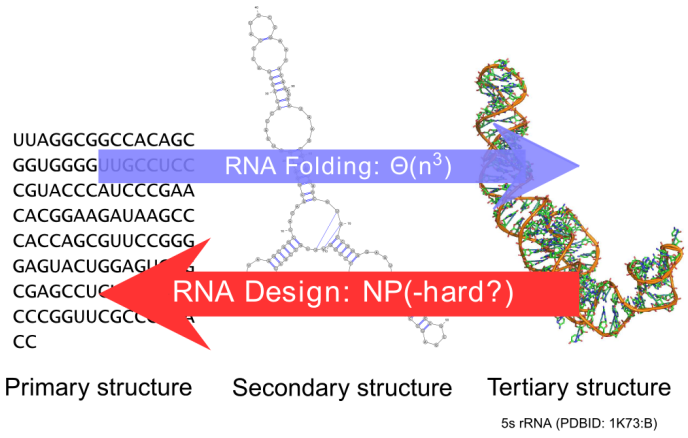


Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

3.1.2. RNA design.

Participants: Alice Heliou, Vincent Le Gallic, Yann Ponty.

Joint project with A. Denise (sc Lri), S. Vialette (Marne-la-Vallée), J. Waldispühl (McGill) and Y. Zhang (Wuhan).

It is a natural pursue 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 our group, we offer a new methodology, based on weighted random generation [33] 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 [39].

3.1.3. Towards 3D modeling of large molecules

Participants: Yann Ponty, Afaf Saaidi, Mireille Regnier.

Joint projects with A. Denise (sc Lri), D. Barth (Versailles), J. Cohen (Paris-Sud), B. Sargueil (Paris V) and Jérome 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 settings directions.

Modeling tasks can also be greatly helped by the availability of experimental data. However, high-resolution techniques such as crystallography or RMN, are notoriously costly in term of time and ressources, leading to the current gap between the amount of available sequences and structural data. As part of a colloboration with B. Sargueil's lab (Faculté de pharmacie, Paris V) funded by the Fondation pour la Recherche medical, we strive to propose a new paradigm for the analysis data produced using a new experimental technique, called SHAPE analysis (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension). This experimental setup produces an accessibility profile associated with the different positions of an RNA, the *shadow* of an RNA. As part of A. Saaidi's PhD, we currently design new algorithmic strategies to infer the secondary structure of RNA from multiple SHAPE experiments performed by experimentalists at Paris V. Those are obtained on mutants, and will be coupled with a fragment-based 3D modeling strategy developed by our partners at McGill.

3.2. Sequences

Participants: Mireille Régnier, Philippe Chassignet, Yann Ponty, Jean-Marc Steyaert, Alice Héliou, Daria Iakovishina, 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 beenestablished 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 [44]. 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 Régnier, Philippe Chassignet, Alice Héliou, Daria Iakovishina.

Besides applications [5] 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. Assessment significance requires a quantitative criteriumsuch as the p-value.

In the recent years, a general scheme of derivation of analytic formula for the pvalue 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 [47]. Recursive equations to compute pvalues 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.

This research area is widened by new problems arising from *de novo* genome assembly or re-assembly.

In [54], 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 [37], and, recently, heuristics have been proposed and implemented [34], [49], [30]. 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 extremly 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 [29], [32]. Recently, it was shown that considering the words which occur in one sequence but do no in another can be used to detect biologically significant events [52]. We have studied the computation of minimal absent words and we have provided new linear implementations [25],[21].

According to the current knowledge, cancer develops as a result of the mutational process of the genomic DNA. In addition to point mutations, cancer genomes often accumulate a significant number of chromosomal rearrangements also called structural variants (SVs). Identifying exact positions and types of these variants may lead to track cancer development or select the most appropriate treatment for the patient. Next Generation Sequencing opens the way to the study of structural variants in the genome, as recently described in [27]. This is the subject of an international collaboration with V. Makeev's lab (IOGENE, Moscow), MAGNOME project-team and V. Boeva (Curie Institute). One goal is to combine two detection techniquesbased either on paired-end mapping abnormalities or on variation of the depth of coverage. A second goal is to develop a model of errors, including a statistical model, that takes into account the quality of data from the different sequencing technologies, their volume and their specificities such as the GC-content or the mappability.

3.2.2. Random generation

Participant: Yann Ponty.

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 [35]: 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. Indeed, in a joint work with A. Denise (LRI), we have done significant and original progress in this area recently [46], [5], including combinatorial models for structures with pseudoknots. Our aim is to combine this paradigm with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [44].

3.3. 3D interaction and structure prediction

Participants: Julie Bernauer, 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 [26], [36] and efficient code are essential [40].

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 is extending the approach to predict protein structures.

3.3.1. Statistical and robotics-inspired models for structure and dynamics

Participants: Julie Bernauer, Amélie Heliou.

Despite being able to correctly model small globular proteins, the computational structural biology community still craves for efficient force fields and scoring functions for prediction but also good sampling and dynamics strategies.

Our current and future efforts towards knowledge-based scoring function and ion location prediction have been described in 3.3.1 .

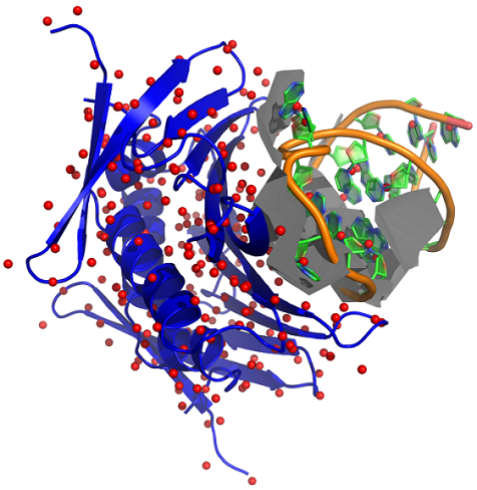


Figure 2. Coarse-grained representation and Voronoi interface model of a PP7 coat protein bound to an RNA hairpin (PDB code 2qux). The Voronoi model captures the features of the interactions such as stacking, even at the coarse-grained level.

Over the last two decades a strong connection between robotics and computational structural biology has emerged, in which internal coordinates of proteins are interpreted as a kinematic linkage with rotatable bonds as joints and corresponding groups of atoms as links [55], [31], [43], [42]. Initially, fragments in proteins limited to tens of residues were modeled as a kinematic linkage, but this approach has been extended to encompass (multi-domain) proteins [41]. For RNA, progress in this direction has been realized as well. A kinematics-based conformational sampling algorithm, KGS, for loops was recently developed [38], but it does not fully utilize the potential of a kinematic model. It breaks and recloses loops using six torsional degrees of freedom, which results in a finite number of solutions. The discrete nature of the solution set in the conformational space makes difficult an optimization of a target function with a gradient descent method. Our methods overcome this limitation by performing a conformational sampling and optimization in a co-dimension 6 subspace. Fragments remain closed, but these methods are limited to proteins. Our objective is to extend the approach proposed in [38], [55] to nucleic acids and protein/nucleic acid complexes with a view towards improving structure determination of nucleic acids and their complexes and in silico docking experiments of protein/RNA complexes. For that purpose, we have developed a generic strategy for differentiable statistical potentials [1], [53] that can be directly integrated in the procedure.

Results from in silico docking experiments will also directly benefit structure determination of complexes which, in turn, will provide structural insights in nucleic acid and protein/nucleic acid complexes. From the small proof-of-concept single chain protein implementation of the KGS strategy, we have developed a robust preliminary implementation that can handle RNA and will be further developed to account for multi-chain , with an extensive computational and biological validation.

BEAGLE Project-Team

3. Research Program

3.1. Introduction

As stated above, the research topics of the BEAGLE Team are centered on the modelisation and simulation of cellular processes. More specifically, we focus on two specific processes that govern cell dynamics and behavior: Evolution and Biophysics. This leads to two main topics: computational cell biology and models for genome evolution.

3.2. Computational Cell Biology

BEAGLE contributes computational models and simulations to the study of cell signaling in prokaryotic and eukaryotic cells, with a special focus on the dynamics of cell signaling both in time and in space. Importantly, our objective here is not so much to produce innovative computer methodologies, but rather to improve our knowledge of the field of cell biology by means of computer methodologies.

This objective is not accessible without a thorough immersion in experimental cell biology. Hence, one specificity of BEAGLE is to be closely associated inside each research project with experimental biology groups. For instance, all the current PhD students implicated in the research projects below have strong interactions with experimenters, most of them conducting experiments themselves in our collaborators' labs. In such a case, the supervision of their PhD is systematically shared between an experimentalist and a theoretician (modeler/computer scientist).

Standard modeling works in cell biochemistry are usually based on mean-field equations, most often referred to as "laws of mass-action". Yet, the derivation of these laws is based on strict assumptions. In particular, the reaction medium must be dilute, perfectly-mixed, three-dimensional and spatially homogeneous and the resulting kinetics are purely deterministic. Many of these assumptions are obviously violated in cells. As already stressed out before, the external membrane or the interior of eukaryotic as well as prokaryotic cells evidence spatial organization at several length scales, so that they must be considered as non-homogeneous media. Moreover, in many case, the small number of molecule copies present in the cell violates the condition for perfect mixing, and more generally, the "law of large numbers" supporting mean-field equations.

When the laws-of-mass-action are invalidated, individual-based models (IBM) appear as the best modeling alternative to evaluate the impact of these specific cellular conditions on the spatial and temporal dynamics of the signaling networks. We develop Individual-Based Models to evaluate the fundamental impact of non-homogeneous space conditions on biochemical diffusion and reaction. More specifically, we focus on the effects of two major sources of non-homogeneity within cells: macromolecular crowding and non-homogeneous diffusion. Macromolecular crowding provides obstacles to the diffusive movement of the signaling molecules, which may in turn have a strong impact on biochemical reactions [42]. In this perspective, we use IBM to renew the interpretation of the experimental literature on this aspect, in particular in the light of the available evidence for anomalous subdiffusion in living cells. Another pertinent source of non-homogeneity is the presence of lipid rafts and/or caveolae in eukaryotic cell membranes that locally alter diffusion. We showed several properties of these diffusion gradients on cells membranes. In addition, combining IBMs and cell biology experiments, we investigate the spatial organization of membrane receptors in plasmic membranes and the impact of these spatial features on the initiation of the signaling networks [46]. More recently, we started to develop IBMs to propose experimentally-verifiable tests able to distinguish between hindered diffusion due to obstacles (macromolecular crowding) and non-homogeneous diffusion (lipid rafts) in experimental data.

The last aspect we tackle concerns the stochasticity of gene expression. Indeed, the stochastic nature of gene expression at the single cell level is now a well established fact [52]. Most modeling works try to explain this stochasticity through the small number of copies of the implicated molecules (transcription factors, in particular). In collaboration with the experimental cell biology group led by Olivier Gandrillon at the Centre de Génétique et de Physiologie Moléculaire et Cellulaire (CGPhyMC, UMR CNRS 5534), Lyon, we study how stochastic gene expression in eukaryotic cells is linked to the physical properties of the cellular medium (e.g., nature of diffusion in the nucleoplasm, promoter accessibility to various molecules, crowding). We have already developed a computer model whose analysis suggests that factors such as chromatin remodeling dynamics have to be accounted for [48]. Other works introduce spatial dimensions in the model, in particular to estimate the role of space in complex (protein+ DNA) formation. Such models should yield useful insights into the sources of stochasticity that are currently not explained by obvious causes (e.g. small copy numbers).

3.3. Models of genome evolution

Classical artificial evolution frameworks lack the basic structure of biological genome (i.e. a double-strand sequence supporting variable size genes separated by variable size intergenic sequences). Yet, if one wants to study how a mutation-selection process is likely (or not) to result in particular biological structures, it is mandatory that the effect of mutation modifies this structure in a realistic way. We have developed an artificial chemistry based on a mathematical formulation of proteins and of the phenotypic traits. In our framework, the digital genome has a structure similar to prokaryotic genomes and a non-trivial genotype-phenotype map. It is a double-stranded genome on which genes are identified using promoter-terminator- like and start-stop-like signal sequences. Each gene is transcribed and translated into an elementary mathematical element (a "protein") and these elements - whatever their number - are combined to compute the phenotype of the organism. The Aevol (Artificial EVOLution) model is based on this framework and is thus able to represent genomes with variable length, gene number and order, and with a variable amount of non-coding sequences (for a complete description of the model, see [59]).

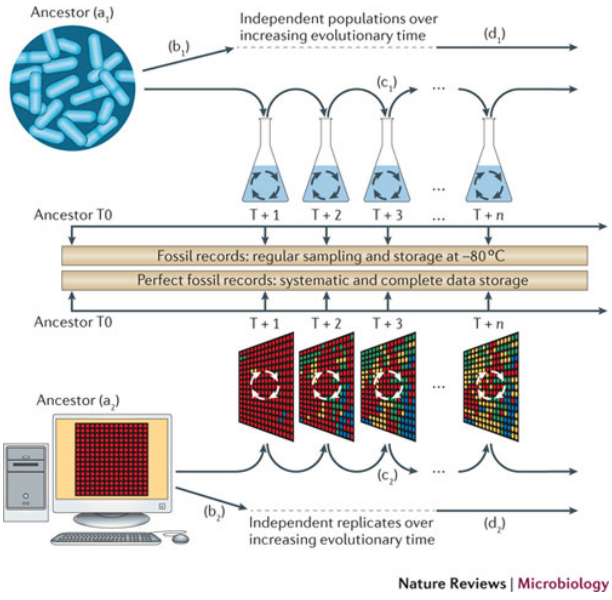


Figure 1. Parallel between experimental evolution and artificial evolution

As a consequence, this model can be used to study how evolutionary pressures like the ones for robustness or evolvability can shape genome structure [60], [57], [58], [67]. Indeed, using this model, we have shown that genome compactness is strongly influenced by indirect selective pressures for robustness and evolvability. By genome compactness, we mean several structural features of genome structure, like gene number, amount of non functional DNA, presence or absence of overlapping genes, presence or absence of operons [60], [57], [68]. More precisely, we have shown that the genome evolves towards a compact structure if the rate of spontaneous mutations and rearrangements is high. As far as gene number is concerned, this effect was known as an error-threshold effect [51]. However, the effect we observed on the amount of non functional DNA was unexpected. We have shown that it can only be understood if rearrangements are taken into account: by promoting large duplications or deletions, non functional DNA can be mutagenic for the genes it surrounds.

We have extended this framework to include genetic regulation (R-Aevol variant of the model). We are now able to study how these pressures also shape the structure and size of the genetic network in our virtual organisms [44], [43], [45]. Using R-Aevol we have been able to show that (i) the model qualitatively reproduces known scaling properties in the gene content of prokaryotic genomes and that (ii) these laws are not due to differences in lifestyles but to differences in the spontaneous rates of mutations and rearrangements [43]. Our approach consists in addressing unsolved questions on Darwinian evolution by designing controlled and repeated evolutionary experiments, either to test the various evolutionary scenarios found in the literature or to propose new ones. Our experience is that "thought experiments" are often misleading: because evolution is a complex process involving long-term and indirect effects (like the indirect selection of robustness and evolvability), it is hard to correctly predict the effect of a factor by mere thinking. The type of models we develop are particularly well suited to provide control experiments or test of null hypotheses for specific evolutionary scenarios. We often find that the scenarios commonly found in the literature may not be necessary, after all, to explain the evolutionary origin of a specific biological feature. No selective cost to genome size was needed to explain the evolution of genome compactness [60], and no difference in lifestyles and environment was needed to explain the complexity of the gene regulatory network [43]. When we unravel such phenomena in the individual-based simulations, we try to build "simpler" mathematical models (using for instance population genetics-like frameworks) to determine the minimal set of ingredients required to produce the effect. Both approaches are complementary: the individual-based model is a more natural tool to interact with biologists, while the mathematical models contain fewer parameters and fewer ad-hoc hypotheses about the cellular chemistry.

At this time, simulating the evolution of large genomes during hundreds of thousands of generation with the Aevol software can take several weeks or even months. It is worse with Raevol, where we not only simulate mutations and selection at the evolutionary timescale, but also simulate the lifetime of the individuals, allowing them to respond to environmental signals. Previous efforts to parallelize and distribute Aevol had yielded limited results due to the lack of dedicated staff on these problems. Since September 2014, we have been improving the performance of (R-)Aevol. Thanks to the ADT Aevol, one and a half full time engineers work on improving Aevol and especially to parallelize it. Moreover, we are working to formalize the numerical computation problems with (R-)Aevol to use state-of-the-art optimization techniques from the HPC community. It ranges from dense and sparse matrix multiplication and their optimizations (such as Tridiagonal matrix algorithm) to using new generation accelerator (Intel Xeon Phi and NVidia GPU). However, our goal is not to become a HPC nor a numerical computation team but to work with well-established teams in these fields, such as through the Joint Laboratory for Extreme-Scale Computing, but also with Inria teams in these fields (e.g. ROMA, Avalon, CORSE, RUNTIME, MESCAL). By doing so, (R-)Aevol simulations will be faster, allowing us to study more parameters in a shorter time. Furthermore, we will also be able to simulate more realistic population sizes, that currently do not fit into the memory of a single computer.

In 2015 we have improved both the quality and the performance of the code. For example, we are currently testing a new representation of the phenotype allowing us to use vector operation. In collaboration with the Avalon team and with the help of a shared internship (Mehdi Ghesh), we have build a benchmark for ordinary differential equation (ODE). This benchmark is based on a representative sample of the ODEs (formalizing the genetic network) found within the R-Aevol model. Thanks to this benchmark, we can compare different ODE solvers and methods. Furthermore, researchers working on ODE solvers and methods could use it to evaluate

the quality of their approach. We are now working with Avalon team on an algorithm that will automatically choose at runtime the best fitting solver and method (from a performance and a quality of results point of view). Through this collaboration, we have also extended the execo experimental engine [0] to support Aevol and R-Aevol. By doing so, we have now a complete automatic workflow to conduct large scale campaign experiments with thousands of different parameters of our model and use the resources of distributed platform (Grid'5000 in our case).

Since 2014, we are also working on a second model of genome evolution.This new model, developed by the team within the Evoevo european Project, encompasses not only the gene regulation network (as Raevol does) but also the metabolic level [36]. It allows us to have a real notion of resources and thus to have more complex ecological interactions between the individuals. To speed up computations, the genomic level is simplified compared to aevol, as a chromosome is modelled as a sequence of genes and regulatory elements and not as a sequence of nucleotides. Both models are thus complementary.

Little has been achieved concerning the validation of these models, and the relevance of the observed evolutionary tendencies for living organisms. Some comparisons have been made between Avida and experimental evolution [61], [55], but the comparison with what happened in a long timescale to life on earth is still missing. It is partly because the reconstruction of ancient genomes from the similarities and differences between extant ones is a difficult computational problem which still misses good solutions for every type of mutations, in particular the ones concerning changes in the genome structure.

There exist good phylogenic models of punctual mutations on sequences [53], which enable the reconstruction of small parts of ancestral sequences, individual genes for example [62]. But models of whole genome evolution, taking into account large scale events like duplications, insertions, deletions, lateral transfer, rearrangements are just being developped [70], [49]. Integrative phylogenetic models, considering both nucleotide subsitutions and genome architectures, like Aevol does, are still missing.

Partial models lead to evolutionary hypotheses on the birth and death of genes [50], on the rearrangements due to duplications [41], [69], on the reasons of variation of genome size [56], [63]. Most of these hypotheses are difficult to test due to the difficulty of *in vivo* evolutionary experiments.

To this aim, we develop evolutionary models to reconstruct the history of organisms from the comparison of their genome, at every scale, from nucleotide substitutions to genome organisation rearrangements. These models include large-scale duplications as well as loss of DNA material, and lateral gene transfers from distant species. In particular we have developed models of evolution by rearrangements [64], methods for reconstructing the organization of ancestral genomes [65], [47], [66], or for detecting lateral gene transfer events [40], [8]. It is complementary with the Aevol development because both the model of artificial evolution and the phylogenetic models we develop emphasize on the architecture of genomes. So we are in a good position to compare artificial and biological data on this point.

We improve the phylogenetic models to reconstruct ancestral genomes, jointly seen as gene contents, orders, organizations, sequences. It requires integrative models of genome evolution, which is desirable not only because they will provide a unifying view on molecular evolution, but also because they will shed light onto the relations between different kinds of mutations, and enable the comparison with artificial experiments from models like Aevol.

Based on this experience, the BEAGLE team contributes individual-based and mathematical models of genome evolution, in silico experiments as well as historical reconstruction on real genomes, to shed light on the evolutionary origin of the complex properties of cells.

[0] Matthieu Imbert, Laurent Pouilloux, Jonathan Rouzaud-Cornabas, Adrien Lèbre, Takahiro Hirofuchi "Using the EXECO toolbox to perform automatic and reproducible cloud experiments" 1st International Workshop on UsiNg and building ClOud Testbeds UNICO, collocated with IEEE CloudCom 2013 2013

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 best understanding, we made the choice to structure them in four items. Even if 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 [113], [78]. Piecewise deterministic processes are non diffusion processes also frequently used in the biological context [60], [77], [69], [63]. Among Markov model, we also developed strong expertise about processes derived from Brownian motion and Stochastic Differential Equations [103], [76], [105]. For instance, knowledge about Brownian or random walk excursions [112], [102] 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 (in higher dimensions of the parameter) fractional fields. The basic aspects of these differential equations are now well understood, mainly thanks to the so-called *rough paths* tools [89], but also invoking the Russo-Vallois integration techniques [104]. The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in [61]. Many generalizations (Gaussian or not) of this model have been recently proposed, see for instance [51] for some Gaussian locally self-similar fields, [82] for some non-Gaussian models, [54] for anisotropic models. Our team has thus contributed [59], [83], [82], [84], [97] and still contributes [53], [55], [54], [85], [73] to this theoretical study: Hölder continuity, fractal dimensions, existence and uniqueness results for differential equations, study of the laws to quote a few examples. On the other hand, because of the observation of longitudinal data for each subject in medicine, we have to care about the random effect due to the subject and to choose adapted models like mixed effect models [86], [50], [18]. In the context of health-care and cost-effectiveness analysis, we are also interested in model of aggregation of different criteria. For this purpose, we develop research about fuzzy binary measures and Choquet integral [72], [90].

3.3. Estimation and control for stochastic processes

When one desires to confront theoretical probabilistic models with real data, statistical tools and control of the dynamics are obviously crucial. As matter of course, we develop inference about stochastic processes that we use for modeling, it is the heart of some of our projects. Control of stochastic processes is also a way to optimise administration (dose, frequency) of therapy.

The monograph [81] is a good reference on the basic estimation techniques for diffusion processes. Some attention has been paid recently to the estimation of the coefficients of fractional or multifractional Brownian motion according to a set of observations. Let us quote for instance the nice surveys [48], [58]. On the other hand, the inference problem for diffusions driven by a fractional Brownian motion has been in its infancy. A good reference on the question is [111], dealing with some very particular families of equations, which do not cover the cases of interest for us. We also recently proposed least-square estimators for these kind of processes [57], [98]. Inference about PDMP is also a recent subject that we want to develop. Our team has a good expertise about inference of the rate jump and the kernel of PDMP [46], [47], [45], [33].

However, there is 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 want to tackle the problem of inference of "Hidden PDMP". It could be also interesting to investigate estimation followed by optimal control for ergodic PDMP. About pharmacokinetics modeling inference, several papers have been reported for the application of system identification techniques. But two issues were ignored in these previous works: presence of timing noise and identification from longitudinal data. In [49], we have proposed a bounded-error estimation algorithm based on interval analysis to solve the parameter estimation problem while taking into consideration uncertainty on observation time instants. Statistical inference from longitudinal data based on mixed effects models [86] can be performed by the *Monolix* software (http://www.lixoft.eu/products/monolix/product-monolix-overview/) developed by the Monolix group chaired by Marc Lavielle and France Mentré, and supported by Inria. We used it to estimate timor growth in [50].

We consider the control of stochastic processes within the framework of Markov Decision Processes [100] and their generalization known as multi-player stochastic games [110], 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 [52]. Regarding complexity, a central topic of research is the analysis of the Policy Iteration algorithm, which has made significant progress in the last years [116], [99], [75], [65], [109], 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 [107], [106], [108], [88] [13], 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 (either unordered or ordered) pairs of these nodes called edges. This type of data is frequently used in various domains of application (in particular in biology) because they provide a mathematical representation of many concepts such as physical or biological structures and networks of relationship in a population. Some attention has recently been focused in the group on modeling and inference for graph data.

Suppose that we know the value of p variables on n subjects (in many applications, we have $n \ll p$). Inference network consists in evaluating the link between two variables knowing the others. [114] gives a very good introduction and many references about network inference and mining. Gaussian Graphical model is a convenient framework to infer network between quantitative variables: there is a edge between two variables if the partial correlation between them is non zero. So the problem is to compute the partial correlations trough the concentration matrix. Many methods are available to infer and test partial correlations in the context $n \ll p$ [114], [92], [68], [71]. However, when dealing with abundance data, because inflated zero data, data are far from gaussian assumption. Some authors work only with the binary "presence-absence" indicator via log-linear [74]. Models for inflated zero variables are not used for network inference and we want to develop them.

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 (critical Galton-Watson trees for example): in this context, the structure of the tree depends on an integer-valued distribution that we estimate from the observation of either only one tree, or a forest. We also work on lossy compression of trees via linear directed acyclic graphs. These methods make us able to compute distances between tree data faster than from the original structures and with a high accuracy. These results are valuable in the context of very large trees arising for instance in biology of plants.

3.5. Regression and machine learning

Regression models or machine learning aim at inferring statistical links between a variable of interest and covariates. It amis also at clustering subjects or variables in set homogeneous sets. In biological study, it is always important to develop adapted learning methods both in the context of "standard" data and also for very massive or online data.

A first approach for regression of quantitative variable is the non-parametric estimation of its cumulative distribution function. Many methods are available to estimate conditional quantiles and test dependencies [96], [79]. Among them we have developped nonparametric estimation trough local analysis via polynomial [66], [67] and we want to study properties of this estimator in order to derive 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 test, thanks to simulation methods, validity of their assumptions. These kind of test are called omnibus test. An omnibus test is an overall test that examines several assumptions together, the most known omnibus test is the one for testing gaussianity (that examines both skewness and kurtosis [62]).

As it concerns the analysis point of high dimensional data, our view on the topic relies on the so-called *French data analysis school*, and more specifically on Factorial Analysis tools. In this context, stochastic approximation is an essential tool (see Lebart's paper [87]), which allows one to approximate eigenvectors in a stepwise manner. A systematic study of Principal Component and Factorial Analysis has then been lead by Monnez in the series of papers [95], [93], [94], in which many aspects of convergences of online processes are analyzed thanks to the stochastic approximation techniques. BiGS aims at performing accurate classification or clustering by taking advantage of the possibility of updating the information "online" using stochastic approximation algorithms [80]. We focus on several incremental procedures for regression and data analysis like linear and logistic regressions and PCA. We also focus the biological context of high-throughput bioassays in which several hundreds or thousands of biological signals are measured for a posterior analysis. The inference of the modeling conclusions from a sample of wells to the whole population requires to account for the inter-individual variability within the modeling procedure. One solution consists in using mixed effects models but up to now no similar approach exists in the field of dynamical system identification. As a consequence, 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 labelled 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 modelled 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 analyse 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 analyse 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 lab on nonribosomal peptides, and has became 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 [54], [32]. 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 [37], [59]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [24].

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 [58]. 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 [7]. 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 [39]. The latest version of Kpax (manuscript in review) 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 [25], 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) [29].

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 [33] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [44]. A different point of view is represented in the Interaction Network Ontology (INO; http://www.ino-ontology.org/ which is a community-driven ontology that is being developed to standardise and integrate data on interaction networks and to support computer-assisted reasoning [60]. 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" [28], 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) [56] functional annotations between standard Uniprot (http://www.uniprot.org/ sequence identifiers and protein structures from the Protein Databank (PDB) [23] 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 have made good progress in this area by developing a recommender-based data mining technique to associate enzyme classification code numbers with Pfam domains using our recently developed EC-DomainMiner program [19].

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 [48], the shape of a protein molecule is represented using polar Fourier series expansions of the form

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

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 [35]

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

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 [49]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [8], [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 [50], [38], 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 dramatically [49][8]. 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 [22], [47], [53]. For example, using our operator notation (in which $\widehat{R}$ and $\widehat{T}$ 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

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

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 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). Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to refine candidate solutions using a more accurate 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 coarse-grained (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 [43], [26], [40], [41]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [9], 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 [21]. 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 [52]. Furthermore, this kind of coarse-graining effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [34]. We are therefore developing a "coarse-grained" scoring function for fast protein-protein docking and multi-component assembly.

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, its 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. Although we do not have precise roadmap to a solution for the multi-component assembly problem, we wish to proceed firstly by putting more emphasis on the single-body terms in the scoring function, 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. Knowledge representation with constraint programming

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. In the last few years, our team has investigated an alternative perspective where each observation 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 need to relax the problem and to study the space of corrections to prior knowledge in order to fit reasonably with observation data. In this case, this issue is modeled as combinatorial (sub)-optimization issues. In both cases, 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 satisfiability problem (in the case of existing solutions) or to the optimization problem (in the case of required corrections to the prior knowledge) [6].

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) [43], [66] 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 [45], and proposed robust corrections [5]. See Fig. 1 for details. The results obtained so far suggest that this approach is compatible with efficiency, scale and expressivity needed by biological systems. Our goal is now to provide **formal models of queries on biological networks** with the focus of integrating dynamical information as explicit logical constraints in the modeling process. This would definitely introduce such logical paradigms as a powerful approach to build and query reconstructed biological systems, in complement to discriminative approaches. 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 whether progresses in the field of constraint logical programming, shown by the performance of ASP-solvers in several recent competitions, are now sufficient to address the complexity of constraint-satisfiability and combinatorial optimization issues explored in systems biology.

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 the models that best fit biological knowledge. 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.

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 opens

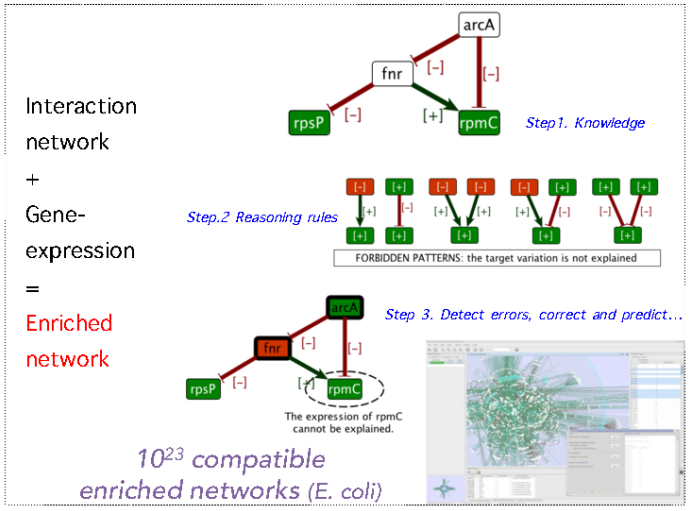


Figure 1. An example of reasoning process in order to identify which expression of non-observed nodes (white nodes) are fixed by partial observations and rules derived from the system dynamics. [6], [5] ***Step 1****. Regulation knowledge is represented as a signed oriented graph. Edge colors stand for regulatory effects (red/green→ inhibition or activation). Vertex colors stand for known gene expression data (red/green→ under or over-expression).* ***Step 2****. Integrity constraints on the whole colored graph come from the necessity to find a consistent explanation of the link between regulation and expression.* ***Step 3****. The model allows both the prediction of values (e.g. for fnr in the figure) and the detection of contradictions (e.g. the expression level of rpmC is inconsistant with the regulation in the graph).*

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.

3.2. Probabilistic and symbolic dynamics

We work on optimization techniques to learn models of the dynamics of a biology systems compatible with a set of quantitative measurements in order to predict its quantitative response at a larger-scale. Our framework mixes mechanistic and probabilistic modeling [2]. The system is modeled by an Event Transition Graph, that is, a **Markovian qualitative description of its dynamics** together with quantitative laws which describe the effect of the dynamic transitions over higher scale quantitative measurements. Then, a few time-series quantitative measurements are provided. Following an ergodic assumption and average case analysis properties, we know that a multiplicative accumulation law on a Markov chain asymptotically follows a log-normal law with explicit parameters [65]. This property can be derived into constraints to describe the set of admissible weighted Markov chains whose asymptotic behavior agrees with the quantitative measures at hand. A precise study of this constrained space via local search optimization emphasizes the most important discrete events that must occur to reproduce the information at hand. These methods have been validated on the *E. coli* regulatory network benchmark. See Figure 2 for illustration. We now plan to apply these techniques to reduced networks representing the main pathways and actors automatically generated from the integrative methods developed in the former section. This requires to improve the range of dynamics that can be modeled by these techniques, as well as the efficiency and scalability of the local search algorithms.

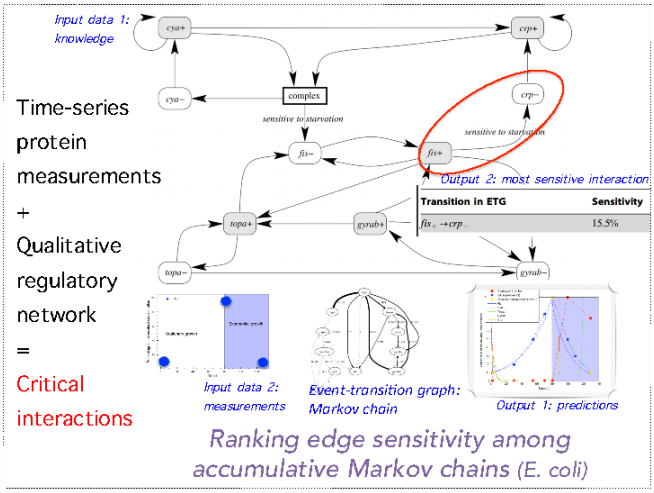


*Figure 2. Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems and identification of key interactions [2]. **Input data** include a qualitative description of the system dynamics at the transcription level (interaction graph) and 3 concentration measurements of the fis protein (population scale). The method computes an **Event-Transition Graph**: interaction frequencies required to predict the population scale behavior as the asymptotic behavior of an accumulation multiplicative law over a Markov chain. Local searches of Markov chains consistent with the observed dynamics and whose asymptotic behavior is consistent with quantitative observations at the population scale. Edge thickness reflects their sensitivity in the search space. It allows to **predict** the Cya protein concentration (red curve) which best fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in E. Coli response to nutritional stress.*

3.3. Modeling sequences with formal grammars

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 [33], we have a strong expertise on learning finite state automata. By introducing a similar fragment merging heuristic approach, we have proposed an algorithm that learns successfully automata modeling families of (non homologous) functional families of proteins [4], leading to a tool named Protomata-learner. As an example, this tool allowed us to properly model the TNF protein family, a difficult task for classical probabilistic-based approaches (see Fig. 3). It was also applied successfully to model important enzymatic families of proteins in cyanobacteria [3]. 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 is 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. We wish to extend our expertise towards inferring Context-Free Grammars including the topological information coming from the structural characterization of active sites.

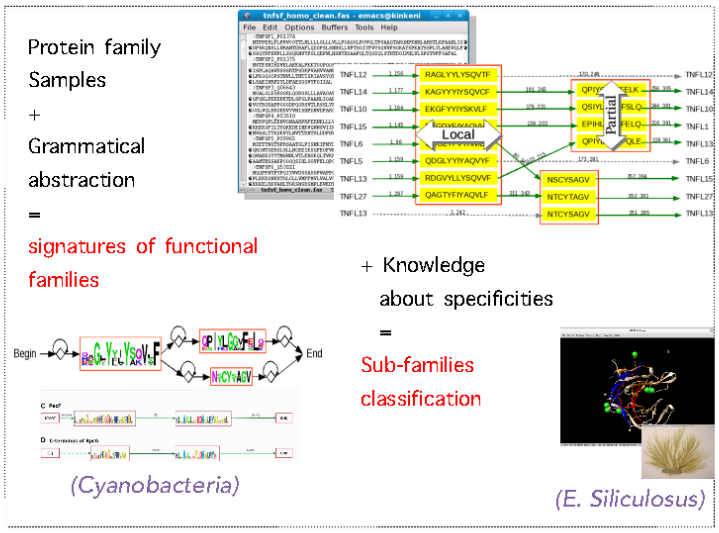


Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences, a partial local alignment is computed and an automaton is inferred, which can be considered as a signature of the family of proteins. This allows searching for new members of the family [3]. Adding further information about the specific properties of proteins within the family allows to exhibit a refined classification.*

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 [70]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [85], [84]. 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. 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 [8]. This tool was used for the recognition of transposable elements in *Arabidopsis thaliana* [88] or for the design of a CRISPR database [10]. See Figure 4 for illustration. Our goal is now to extend the framework of STAN. 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. In this direction, we are working on Logol [1], a language and framework based on a systematic introduction of constraints on string variables.

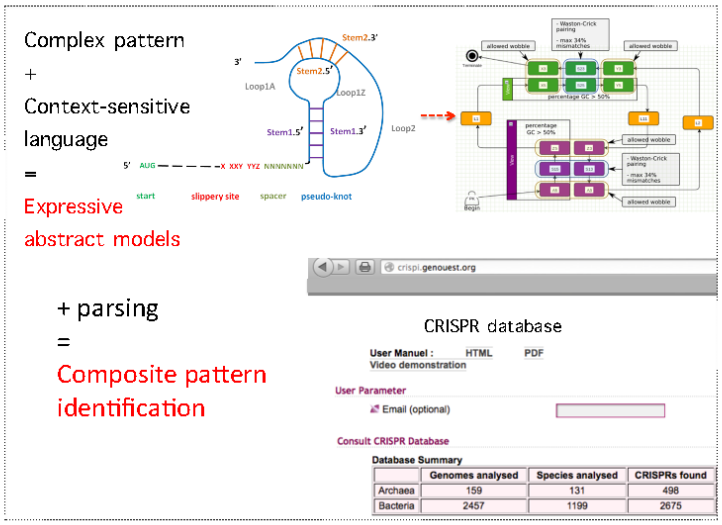


Figure 4. Graphical modeling of a pseudo-knot (RNA structure) based on the expressivity of String Variable Grammars used in the Logol framework. Combined with parsers, this leads to composite pattern identification such as CRISPR [79].

3.4. Symbolic methods for model space exploration: Ontologies and Formal Concepts Analysis

All methods presented in the previous section usually result in pools of candidates which equivalently explain the data and knowlegde. 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. We rely on two classes of symbolic technics to this end: Semantic Web technologies and Formal Concept Analysis (FCA). They both aim at the formalization and management of knowledge, that is, the explicitation of relations occuring in structered data. These technics are complementary: The production of

relevant concepts in FCA highly depends on the availability of semantic annotations using a controlled set of terms and conversely, building ontologies is a complex process that can be made much easier with FCA.

3.4.1. Semantic web for life sciences

Life sciences are intrinsically complicated and complex. Until a few years ago, both the scarcity of available information and the limited processing power imposed the double constraints that work had to be performed on fragmented areas (either precise but narrow or broad but shallow) as well as using simplifying hypotheses [52]. The recent joint evolution of data acquisition capabilities in the biomedical field, and of the methods and infrastructures supporting data analysis (grids, the Internet...) resulted in an explosion of data production in complementary domains (*omics, phenotypes and traits, pathologies, micro and macro environment...) [52], [56], [47]. This "data deluge" is the life-science version of the more general "big data" phenomenon, with the specificities that the proportion of generated data is much higher, and that these data are highly connected [86]. In addition to the breakthrough in each of these domains, major efforts have been undertaken notably in Systems Biology for developing the links between them. **The bottleneck that once was data scarcity now lies in the lack of adequate methods supporting data integration, processing and analysis**. Each of these steps typically hinges on domain knowledge, which is why it resists 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.

The knowledge we are focusing on is mostly symbolic, as opposed to other kinds of biomedical knowledge (probabilistic, related to chemical kinetics, 3D models of anatomical entities or 4D models of processes...). It should typically support generalization, association and deduction. There is a long tradition of works in order to come up with an explicit and formal representation of this knowledge that would support automatic processing.

This line of work resulted in the now widespread acceptance of ontologies [87], [59] to represent the biomedical entities, their properties and the relations between these entities. Bard et al. defined **ontologies** as "formal representations of knowledge in which the essential terms are combined with structuring rules that describe the relationships between the terms" [44]. Ontologies range from fairly simple hierarchies to semantically-rich organization supporting complex reasoning [59]. Ontologies are now a well established field [59], [54] that evolved from concept representation [83].

The emergence of ontologies in biomedical informatics and bioinformatics happened in parallel with the development of the **Semantic Web** in the computer science community [81], [83]. The Semantic Web is an extension of the current Web that provides an infrastructure integrating data and ontologies in order to support unified reasoning.

Life sciences are a great application domain for the Semantic Web [57], [76], [46]. Semantic Web technologies have become an integral part of translational medicine and translational bioinformatics [47], [58]. The Linked Data initiative [51] and particularly the Linked Open Data project promotes the integration of data sources in machine-processable formats compatible with the Semantic Web. Figure 5 shows the importance of life sciences. In the past few years, this proved instrumental for addressing the problem of data integration [68], [72].

We are working on the integration of Semantic Web resources with our data analysis methods in order to take existing biological knowledge into account.

3.4.2. Formal Concept Analysis

Initially developed in the community of set and order theorists, algebraists and discrete mathematicians, formal concept analysis aims at the development of conceptual structures which can be logically activated for the formation of judgments and conclusions [90]. In its most simple form, one considers a binary relation between a set of objects O and a set of attributes A. The derivation operator ' associates to each subset U of O (resp. V of A) the subset of elements in A (resp. O) related to all elements in U (resp. V). A formal concept is characterized by an extension (subset of O, individuals belonging to the concept) and an intension (subset of A, properties applying to all objects in the extension), such that the two subsets are stable sets under the

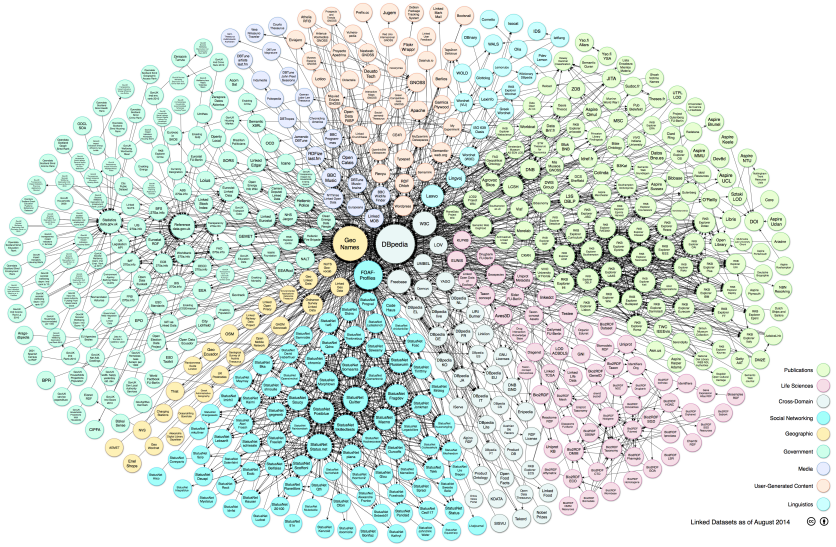


Figure 5. Linked Open Data cloud in August 2014. Nodes are resources. Edges are cross-references between resources. Life science resources constitute the purple portion in the lower right corner.

double derivation relation ". Concepts are related within a lattice structure (Galois connection) by subconcept-superconcept relations, and this allows to draw causality relations between attribute subsets.

It is used in various domains managing structered data such as knowledge processing, information retrieval or classification [73]. We study the issues raised by its application in bioinformatics. Among others, it has been used to derive phylogenetic relations among groups of organisms [71], 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 [61]. 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. This has been used for example for the drug repositioning issue [62], where the completion of concepts is used as a support for the prediction of new relations in a drug-target-disease network and ultimately the assignment of drugs to new diseases. 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 [91]. The general idea consists to look for a tradeoff between the simplicity of the conceptual representation and the need to manage exceptions. We are also interested in using ontologies to help this process or to help ontology refinement using concept analysis [74], [50], [78].

Networks are widely used in bioinformatics for the integration of multiple sources of data inside a common model and this leads to very large networks (protein/protein interactions, signaling or regulation network, metabolic network...). 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 [80]. In fact, concepts are maximal bicliques and we are interested in developing the power graph idea in the framework of concept analysis.

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 substracting 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 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. The first four axes thus fall in the first category, and the fifth one in the second. As concerns the first four axes, the model organisms or systems chosen will be those studied by the biologists among our permanent members or among our close collaborators. Currently these include the following cases:

- Arthropods, notably insects, and their parasites;
- Symbiont-harbouring trypanosomatids and trypanosomas 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: 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 synthetises (for instance, ethanol) but not in enough quantity, or to produce a metabolite it normally is not able to synthetise. 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 5: 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. Introduction

Based on the overall objectives, the research program of GenScale is structured into four research axes as described below. The first three axes include pure computer science aspects, such as the development of advanced data structures and/or the design of new optimized algorithms; they also include strong partnerships with life science actors to validate the methodologies that are developed. The fourth axis can be seen as a transversal one. It addresses efficient parallel implementations of our methods on standard processors, cluster systems, or accelerators such as GPU.

3.2. Axis 1: HTS data processing

The raw information delivered by NGS (Next Generation Sequencing) technologies represents billions of short DNA fragments. An efficient structuration of this mass of data is the de-Bruijn graph that is used for a large panel of problems dealing with high throughput genomic data processing. The challenge, here, is to represent this graph into memory. An efficient way is to use probabilistic data structures, such as Bloom filters but they generate false positives that introduce noise and may lead to errors. Our approach is to enhance this basic data structure with extra information to provide exact answers, while keeping a minimal memory occupancy [2], [3].

Based on this central data structure, a large panel of HTS algorithms can be designed: read compression, read correction, genome assembly, detection of SNPs (Single Nucleotide Polymorphism) or detection of other variants such as inversion, transposition, etc. [8], [11], [9]. The use of this compact structure guarantees software with very low memory footprint that can be executed on many standard-computing resources.

In the full assembly process, an open problem due to the structure complexity of many genomes is the scaffolding step that consists in reordering contigs along the chromosomes. This treatment can be formulated as a combinatorial optimization problem exploiting the upcoming new sequencing technologies based on long reads.

3.3. Axis 2: Sequence comparison

Comparing genomic sequences (DNA, RNA, protein) is a basic bioinformatics task. Powerful heuristics (such as the seed-extend heuristic used in the well-known BLAST software) have been proposed to limit the computation time. The underlying data structures are based on seed indexes allowing a drastic reduction of the search space. However, due to the increasing flux of genomic sequences, this treatment tends to increase and becomes a critical section, especialy in metagenomic projects where hundred of millions of reads must be compared to large genomic banks for taxonomic of functional assignation.

Our research follows mainly two directions. The first one revisits the seed-extend heuristic in the context of the bank-to-bank comparison problem. It requires new data structures to better classify the genomic information, and new algorithmic methods to navigate through this mass of data [7], [10]. The second one addresses metagenomic challenges that have to extract relevant knowledge from Tera bytes of data. In that case, the notion of sequence similarity itself is redefined in order to work on objects that are much simpler than the standard alignment score, and that are better suited for large-scale computation. Raw information (reads) is first reduced to k-mers from which high speed and parallel algorithms compute approximate similarities based on a well defined statistical model [4].

3.4. Axis 3: Protein 3D structure

The three-dimensional (3D) structure of proteins tends to be evolutionarily better preserved during evolution than its sequence. Finding structural similarities between proteins gives deep insights into whether these proteins share a common function or whether they are evolutionarily related. Structural similarity between two proteins is usually defined by two functions – a one to- one mapping (also called alignment) between two subchains of their 3D representations and a specific scoring function that assesses the alignment quality. The structural alignment problem is to find the mapping that is optimal with respect to the scoring function. Protein structures can be represented as graphs, and the problem reduces to various combinatorial optimization problems that can be formulated in this framework: for example finding the maximum weighted path [1] or finding the maximum cardinality clique/pseudo-clique [5] [18].

In most cases, however, suitable conformations for a given protein are unknown. To support this statement, we point out that the number of deposited protein conformations on the Protein Data Bank (PDB [0]) recently reached the threshold of 110,000 entries, while the UniProtKB/TrEMBL [0] database contains more than 50 million sequence entries, all of them potentially capable for coding for a new protein. In this context, distance geometry provides powerful methods and algorithms for the identification of protein conformations from Nuclear Magnetic Resonance (NMR) data, which basically consist of a distance list concerning atom pairs of the protein[6]. We are working on the discretization of the distance geometry, so that its search space becomes discrete (and finite!), for making it possible to perform an exhaustive exploration of the solution set.

3.5. Axis 4: Parallelism

Together with the design of new data structures and new algorithms, our research program aims to propose efficient hardware implementation. Even if not explicitly mentioned in the three previous axes, we have constantly in mind to exploit the parallelism of current processors. Practically, and depending of the nature of the computation to perform, three levels of parallelism are addressed: the use of vector instructions of today processors, the multithreading offered by multi-core systems, and the cluster (or cloud) infrastructures.

Consequent bioinformatics treatments, from the processing of raw HTS data to high-level analysis, are generally performed within a workflow environment and executed on cluster systems. Automating the parallelization of such treatments directly from a graphical capture of the workflow is a necessity for end-users that are generally not expert in parallelism. The challenge here is to hide, as much as possible, the different transformations to go from a high level workflow description to an efficient parallel execution that exploits both task-level and data-level parallelism[25].

Another research activity of this axe is the design of parallel algorithms targeting hardware accelerators, especially GPU boards (Graphical Processing Unit). These devices now offer a high-level programming environment to access the hundred of processors available on a single chip [20]. A few bioinformatics treatments, such as ones that exhibit good computational regularity, can highly benefit from the computing power of this technology.

[0] http://www.rcsb.org/
[0] http://www.ebi.ac.uk/uniprot/TrEMBLstats

IBIS Project-Team

3. Research Program

3.1. Analysis of qualitative dynamics of gene regulatory networks

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

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.

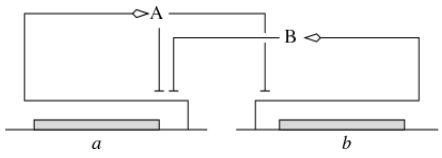
A variety of qualitative modeling formalisms have been introduced over the past decades. Boolean or logical models, which describe gene regulatory and signalling networks as discrete-time finite-state transition systems, are probably most widely used. The dynamics of these systems are governed by logical 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 (hyperrectangular) 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, Stéphan Lacour, Michel Page, Corinne Pinel, Delphine Ropers, Valentin Zulkower.

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.



$$\dot{x}_a = \kappa_a \, s^-(x_a, \theta_a^2) \, s^-(x_b, \theta_b) - \gamma_a \, x_a$$

$$\dot{x}_b = \kappa_b \, s^-(x_a, \theta_a^1) - \gamma_b \, x_b$$

$$s^+(x, \theta) = \begin{cases} 1, & \text{if } x > \theta \\ 0, & \text{if } x < \theta \end{cases}$$

$$s^-(x, \theta) = 1 - s^+(x, \theta)$$

(a) (b)

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.

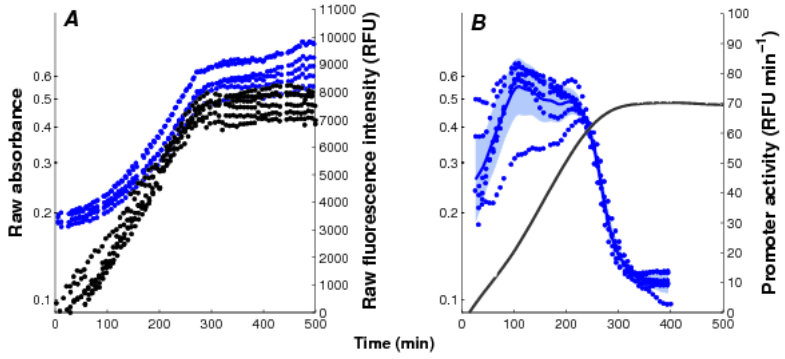


Figure 3. Monitoring of bacterial gene expression in vivo using fluorescent reporter genes (Stefan et al., PLoS Computational Biology, 11(1):e1004028, 2015). The plots show the primary data obtained in a kinetic experiment with E. coli cells, focusing on the expression of the motility gene tar in a mutant background. A: Absorbance (•, black) and fluorescence (•, blue) data, corrected for background intensities, obtained with the ΔcpxR 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 ± 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 (Sections 5.4 and 5.3). 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 (Switzerland) and the University of Pavia (Italy). 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, Johannes Geiselmann, Stéphan Lacour, Yves Markowicz, Manon Morin, Michel Page, Corinne Pinel, Stéphane Pinhal, Delphine Ropers [Correspondent], Valentin Zulkower.

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 hetereogeneous 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: Cindy Gomez Balderas, Eugenio Cinquemani, Johannes Geiselmann [Correspondent], Nils Giordano, Hidde de Jong, Stéphan Lacour, Delphine Ropers, Alberto Soria-Lopéz.

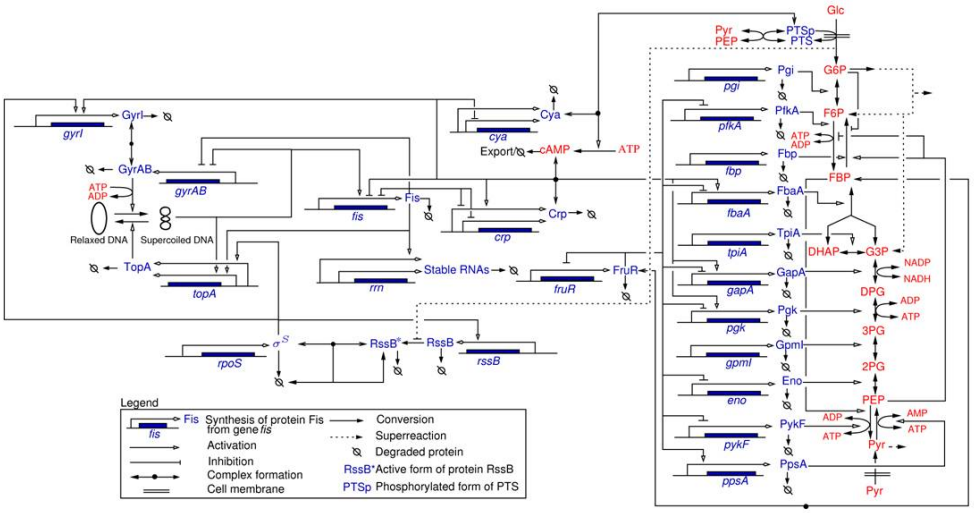


Figure 4. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in E. coli (Baldazzi et al., PLoS Computational Biology, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the 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 envrionmental 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 Université Paris Descartes.

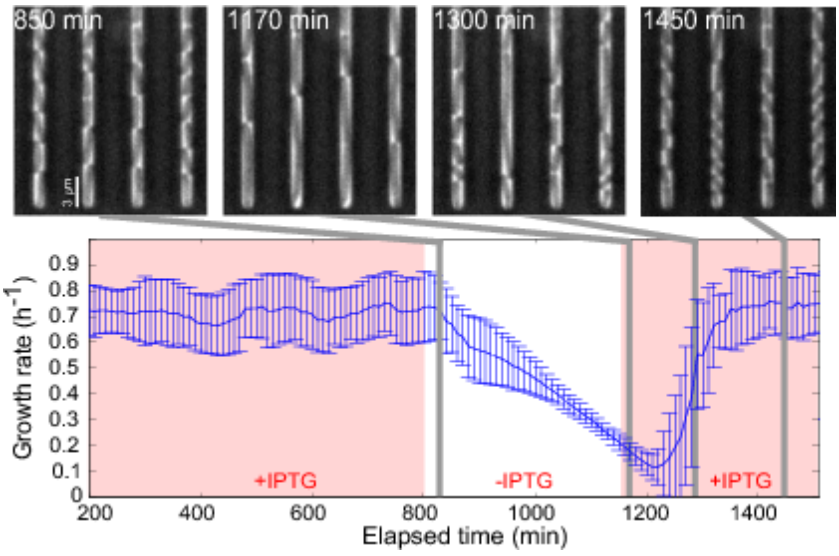


Figure 5. Growth arrest by external control of the gene expression machinery (Izard, Gomez Balderas et al., Molecular Systems Biology, 11:840, 2015). An E. coli strain in which an essential component of the gene expression machinery, the $\beta\beta'$ subunits of RNA polymerase, was put under the control 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 microfluidic 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. Modeling of 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). In addition to the widely-used ordinary differential equation modeling framework, stochastic modeling frameworks, such as chemical master equations, and statistic modeling frameworks, such as ensemble models, are increasingly popular, since they enable to capture biological variability.

In all cases, dedicated mathematical and computational approaches are needed for the analysis of the 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 models, and (stochastic approximations of) the expectation maximization algorithm for the identification of mixed-effects models.

3.3. 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.

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 recently that this platform enables controlling the level of a fluorescent protein in cells with unprecedented accuracy and for many cell generations [0].

3.4. Chemical Reaction Network 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, which can be predicted by static analyzers without 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 systems [0], the inference of reaction systems from ODEs [7], the computation of structural invariants by constraint programming techniques, and the analysis of model reductions by subgraph epimorphisms [2], now provide solid ground for developing static analyzers, using them on a large scale in systems biology, and elucidating modules.

3.5. 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 developped for quantitative models can be summarized by the following identifications :

biological model = transition system,

biological property = temporal logic formula,

model validation = model-checking,

model inference = constraint solving.

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 optimization. This line of research continues with the development of patterns with efficient solvers and their generalization to handle stochastic effects.

3.6. Constraint solving and optimization

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

[0] Jannis Uhlendorf, Agnés 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.

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

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

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

up to 100 parameters, for building challenging quantitative models, gaining model-based insights, revisiting admitted assumptions and contributing to biological knowledge [0]

[0]Domitille 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 genetical 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 experimentations 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. Distances and pattern recognition

Diversity may be understood as a set of dissimilarities between objects. The underlying mathematical construction is the notion of distance. Knowing a set of objects, on the condition that pairwise distances can be measured, it is possible to build a Euclidan 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. It is still true that the reference for recognizing patterns or shapes is the human eye. One objective of our project is to narrow the gap between the story that a human eye can read, and the story that an algorithm can tell. Several directions will be explored. First, it is necessary to master dimension reduction, mainly classical algebraic tools (PCA, NGS, Isomap, eigenmaps, etc ...), and collaborate with experts in efficient methods in spectral methods. Second, a neighborhood in a point cloud naturally leads to graphs describing the neighborhood networks. There is a natural link between modular structures in distance arrays and communities on graphs. Third, points defined by DNA sequences (for example) are samples of diversity. Dimension reduction may show that they live on a given manifold. This leads to geometry (differential or Riemanian geometry). Knowing some properties of the manifold can inform us about the constraints on the space where the measured individuals live. The connection between Riemannian geometry and graphs, where weighted graphs are seen as meshes embedded in a manifold, is currently an active field of reasearch [23], [22].

To resolve these objectives computationally will require investment in research directions in computational geometry (such as convex hulls of high-dimension sets of points), on circumventing the curse of dimensionality, and on linking distance geometry with convex optimization procedures through matrix completion. None of these questions is trivial: most recent work has focused on two or three dimensions, for example for image analysis or for reconstruction of protein conformation from local distances between atoms. The methodological goal is to extend these approaches to higher dimension spaces.

3.2. 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 [15]. 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* [11] 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 certains kinds of systems in biotechnology [14], [16] and medicine [13]. 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

Many live-cell fluorescence imaging experiments are limited in time to prevent phototoxicity and photobleaching. The amount of light and time required to observe entire cell divisions can generate biological artifacts. In order to produce images compatible with the dynamic processes in living cells as seen in video-microscopy, we study the potential of denoising, superresolution, tracking, and motion analysis methods in the Bayesian and the robust statistics framework to extract information and to improve image resolution while preserving cell integrity.

In this area, we have already demonstrated that image denoising allows images to be taken more frequently or over a longer period of time [5]. The major advantage is to preserve cell integrity over time since spatio-temporal information can be restored using computational methods [8], [2], [9], [4]. This idea has been successfully applied to wide-field, spinning-disk confocal microscopy [1], TIRF [40], fast live imaging and 3D-PALM using the OMX system in collaboration with J. Sedat and M. Gustafsson at UCSF [5]. The corresponding ND-SAFIR denoiser software (see Section 6.5) has been licensed to a large set of laboratories over the world. New information restoration and image denoising methods are currently investigated to make SIM imaging compatible with the imaging of molecular dynamics in live cells. Unlike other optical sub-diffraction limited techniques (e.g. STED [51], PALM [41]) SIM has the strong advantage of versatility when considering the photo-physical properties of the fluorescent probes [48]. Such developments are also required to be compatible with "high-throughput microscopy" since several hundreds of cells are observed at the same time and the exposure times are typically reduced.

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

3.2.1. Motion analysis and tracking

The main challenge is to detect and track xFP tags with high precision in movies representing several Giga-Bytes of image data. The data are most often collected and processed automatically to generate information on partial or complete trajectories. Accordingly, we address both the methodological and computational issues involved in object detection and multiple objects tracking in order to better quantify motion in cell biology. Classical tracking methods have limitations as the number of objects and clutter increase. It is necessary to correctly associate measurements with tracked objects, i.e. to solve the difficult data association problem [58]. Data association even combined with sophisticated particle filtering techniques [61] or matching techniques [59] is problematic when tracking several hundreds of similar objects with variable velocities. Developing new optical flow and robust tracking methods and models in this area is then very stimulating since the problems we have to solve are really challenging and new for applied mathematics. In motion analysis, the goal is to formulate the problem of optical flow estimations in ways that take physical causes of brightness constancy violations into account [44], [49]. The interpretation of computed flow fields enables to provide spatio-temporal signatures of particular dynamic processes (e.g. Brownian and directed motion) and could help to complete the traffic modelling.

3.2.2. Event detection and motion classification

Protein complexes in living cells undergo multiple states of local concentration or dissociation, sometimes associated with diffusion processes. These events can be observed at the plasma membrane with TIRF microscopy. The difficulty arises when it becomes necessary to distinguish continuous motions due to trafficking from sudden events due to molecule concentrations or their dissociations. Typically, plasma membrane vesicle docking, membrane coat constitution or vesicle endocytosis are related to these issues.

Several approaches can be considered for the automatic detection of appearing and vanishing particles (or spots) in wide-field and TIRF microscopy images. Ideally this could be performed by tracking all the vesicles contained in the cell [61], [46]. Among the methods proposed to detect particles in microscopy images [62], [60], none is dedicated to the detection of a small number of particles appearing or disappearing suddenly between two time steps. Our way of handling small blob appearances/dis-appearances originates from the observation that two successive images are redundant and that occlusions correspond to blobs in one image which cannot be reconstructed from the other image [1] (see also [42]). Furthermore, recognizing dynamic protein behaviors in live cell fluorescence microscopy is of paramount importance to understand cell mechanisms. In our studies, it is challenging to classify intermingled dynamics of vesicular movements, docking/tethering, and ultimately, plasma membrane fusion of vesicles that leads to membrane diffusion or exocytosis of cargo proteins. Our aim is then to model, detect, estimate and classify subcellular dynamic events in TIRF microscopy image sequences. We investigate methods that exploits space-time information extracted from a couple of successive images to classify several types of motion (directed, diffusive (or Brownian) and confined motion) or compound motion.

3.3. From models to image data: simulation and modelling of membrane transport

Mathematical biology is a field in expansion, which has evolved into various branches and paradigms to address problems at various scales ranging from ecology to molecular structures. Nowadays, system biology [52], [64] aims at modelling systems as a whole in an integrative perspective instead of focusing on independent biophysical processes. One of the goals of these approaches is the cell in silico as investigated at Harvard Medical School (http://vcp.med.harvard.edu/) or the VCell of the University of Connecticut Health Center (http://www.nrcam.uchc.edu/). Previous simulation-based methods have been investigated to explain the spatial organization of microtubules [53] but the method is not integrative and a single scale is used to describe the visual patterns. In this line of work, we propose several contributions to combine imaging, traffic and membrane transport modelling in cell biology.

In this area, we focus on the analysis of transport intermediates (vesicles) that deliver cellular components to appropriate places within cells. We have already investigated the concept of Network Tomography (NT) [63] mainly developed for internet traffic estimation. The idea is to determine mean traffic intensities based on statistics accumulated over a period of time. The measurements are usually the number of vesicles detected at each destination region receiver. The NT concept has been investigated also for simulation [3] since it can be used to statistically mimic the contents of real traffic image sequences. In the future, we plan to incorporate more prior knowledge on dynamics to improve representation. An important challenge is to correlate stochastic, dynamical, one-dimensional *in silico* models studied at the nano-scale in biophysics, to 3D images acquired in vivo at the scale of few hundred nanometers.

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. Theses 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 and generalized linear mixed models for longitudinal count data.
  - 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, semi-Markov switching linear mixed models and semi-Markov switching generalized linear 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. General aim

The overall aim of our project is to design new computational and mathematical approaches for studying brain structure (based on anatomical and diffusion MRI) and functional connectivity (based on EEG, MEG and intracerebral recordings). The goal is to transform raw unstructured images and signals into formalized, operational models such as geometric models of brain structures, statistical population models, and graph-theoretic models of brain connectivity. This general endeavor is addressed within the three following main objectives.

3.2. Modeling brain structure: from imaging to geometric models

Structural MRI (anatomical or diffusion-weighted) allows studying in vivo the anatomical architecture of the brain. Thanks to the constant advance of these imaging techniques, it is now possible to visualize various anatomical structures and lesions with a high spatial resolution. Computational neuroanatomy aims at building models of the structure of the human brain, based on MRI data. This general endeavor requires addressing the following methodological issues: i) the extraction of geometrical objects (anatomical structures, lesions, white matter tracks...) from anatomical and diffusion-weighted MRI; ii) the design of a coherent mathematical framework to model anatomical shapes and compare them across individuals. Within this context, we pursue the following objectives.

First, we aim to develop new methods to segment anatomical structures and lesions. We are most specifically interested in the hippocampus, a structure playing a crucial role in Alzheimer's disease, and in lesions of vascular origin (such as white matter hyperintensities and microbleeds). We pay particular attention to the robustness of the approaches with respect to normal and pathological anatomical variability and with respect to differences in acquisition protocols, for application to multicenter studies. We dedicate specific efforts to the validation on large populations of coming from patients data acquired in multiple centers.

Then, we develop approaches to estimate templates from populations and compare anatomical shapes, based on a diffeomorphic deformation framework and matching of distributions. These methods allow the estimation of a prototype configuration (called template) that is representative of a collection of anatomical data. The matching of this template to each observation gives a characterization of the anatomical variability within the population, which is used to define statistics. In particular, we aim to design approaches that can integrate multiple objects and modalities, across different spatial scales.

3.3. Modeling dynamical 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 and perceptual tasks. Indeed, mounting evidence suggests that impairment of such mechanisms might be the first step of a chain of events triggering several neurological disorders, such as the abnormal synchronization of epileptic activities. Naturally, neuroimaging studies investigating functional connectivity in the brain have become increasingly prevalent.

Our team develops a framework for the characterization of brain connectivity patterns, based on connectivity descriptors from the theory of complex networks. The description of the connectivity structure of neural networks is able to characterize for instance, the configuration of links associated with rapid/abnormal synchronization and information transfer, wiring costs, resilience to certain types of damage, as well as the balance between local processing and global integration. Furthermore, we propose to extend this framework to study the reconfiguration of networks over time. Indeed, neurophysiological data are often gathered from longitudinal recording sessions of the same subject to study the adaptive reconfiguration of brain connectivity. Finally, connectivity networks are usually extracted from different brain imaging modalities (MEG, EEG, fMRI or DTI) separately. Methods for combining the information carried by these different networks are still missing. We thus propose to combine connectivity patterns extracted from each modality for a more comprehensive characterization of networks.

3.4. Methodologies for large-scale datasets

Until recently, neuroimaging studies were often restricted to series of about 20-30 patients. As a result, such studies had a limited statistical power and could not adequately model the variability of populations. Thanks to wider accessibility of neuroimaging devices and important public and private funding, large-scale studies including several hundreds of patients have emerged in the past years. In the field of Alzheimer's disease (AD) for instance, one can cite the Alzheimer's Disease Neuroimaging Initiative (ADNI) including about 800 subjects (patients with AD or mild cognitive impairment (MCI) and healthy controls) or the French cohort MEMENTO including about 2000 subjects with memory complaint. These are most often multicenter studies in which patients are recruited over different centers and images acquired on different scanners. Moreover, cohort studies include a longitudinal component: for each subject, multiple images are acquired at different time points. Finally, such datasets often include multimodal data: neuroimaging, clinical data, cognitive tests and genomics data. These datasets are complex, high-dimensional and often heterogeneous, and thus require the development of new methodologies to be fully exploited.

In this context, our objectives are:

- to develop methodologies to acquire and standardize multicenter neuroimaging data;
- to develop imaging biomarkers based on machine learning and longitudinal models;
- to design multimodal analysis approaches for bridging anatomical models and genomics.

The first two aspects focus on neuroimaging and are tightly linked with the CATI project. The last one builds on our previous expertise in morphometry and machine learning, but aims at opening new research avenues combining imaging and "omics" data. This is developed in strong collaboration with the new biostatistics/bioinformatics platform of the IHU-A-ICM.

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 [90]. 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 [81], [79]. Regarding target applications, a good review of the state of the art can be found in the book *Computer Integrated Surgery* [77], in N. Ayache's article [85] and in recent review articles [86], [90]. The scientific journals *Medical Image Analysis* [72], *Transactions on Medical Imaging* [78], and *Computer Assisted Surgery* [80] 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 [75], [76]) and ISBI'2010 (Int. Symp. on Biomedical Imaging [74]) 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 [91], [105]. 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 [108], or to measure some functional properties of the heart from dynamic sequences of Magnetic Resonance [84], Ultrasound or Nuclear Medicine images [92].

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 multi-modal images [0] for each single patient.

[0]Multisequence (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.

[0]Multimodal 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 [107]). Significant efforts have since been made to develop a theory for statistical shape analysis (one can refer to [89] for a good summary, and to the special issue of Neuroimage [106] 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 [102]:

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.

[0] The 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 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 [103], [97], [87], [104], [93]). 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 [88]:

- the first level is mainly geometrical, and addresses the construction of a digital description of the anatomy [82], essentially acquired from medical imagery;
- the second level is physical, involving mainly the biomechanical modeling of various tissues, organs, vessels, muscles and bone structures [95];
- the third level is physiological, involving the modeling of the functions of the major organ systems [96] (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 [73].

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 [101]). 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 [96], [97]). 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 [59], Merboldt et al [63] and Taylor et al [71]. 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) [70], [48]. Historically, the first model in dMRI is the 2nd order diffusion tensor (DTI) [47], [46] 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 [60], [55]. 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 drived 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 [62], [9] and [61]).

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) [74] and High Angular Resolution Diffusion Imaging (HARDI) methods such as Q-ball imaging [72] and other multi-tensors and compartment models [68], [69], [40], [39], [67] 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 [54]. They are at the core of our probabilistic and deterministic tractography algorithms devised to best exploit the full distribution of the fiber ODF (see [51], [4] and [52],[5]).

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 [76], [77] or 4th order Tensor Model [45]. For more details, we refer the reader to our articles in [56], [68] where we review HOT models and to our articles

in [8], 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) [6].

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 [39], [40]. 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 [3], [2], these imaging modalities raise a large amount of mathematical and computational challenges at the core of the research we develop at ATHENA [58], [68].

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 [3]. 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 [64].

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 [64], [49], [50]. 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 [64], [49], [50].

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 download [0], 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 [57]

[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 [41], AxCaliber [42] and NODDI [75] 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 [44], [43], the Solid Harmonic (SoH) basis [53], the Simple Harmonic Oscillator based Reconstruction and Estimation (SHORE) basis [65] and more recent Mean Apparent Propagator or MAP-MRI basis [66].

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.

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 superseeded 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. Using additional a priori anatomical 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 ODYSSÉE/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 [7], [10] and means to calibrate them [73] 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.

DEMAR Project-Team

3. Research Program

3.1. Modelling and identification of the sensory-motor system

Participants: Mitsuhiro Hayashibe, Christine Azevedo Coste, David Guiraud.

The literature on muscle modelling is vast, but most of research works focus separately on the microscopic and on the macroscopic muscle's functional behaviours. The most widely used microscopic model of muscle contraction was proposed by Huxley in 1957. The Hill-Maxwell macroscopic model was derived from the original model introduced by A.V. Hill in 1938. We may mention the most recent developments including Zahalak's work introducing the distribution moment model that represents a formal mathematical approximation at the sarcomere level of the Huxley cross-bridges model and the works by Bestel and Sorine (2001) who proposed an explanation of the beating of the cardiac muscle by a chemical control input connected to the calcium dynamics in the muscle cells, that stimulates the contractile elements of the model. With respect to this literature, our contributions are mostly linked with the model of the contractile element, through the introduction of the recruitment at the fibre scale formalizing the link between FES parameters, recruitment and Calcium signal path. The resulting controlled model is able to reproduce both short term (twitch) and long term (tetanus) responses. It also matches some of the main properties of the dynamic behaviour of muscles, such as the Hill force-velocity relationship or the instantaneous stiffness of the Mirsky-Parmley model. About integrated functions modelling such as spinal cord reflex loops or central pattern generator, much less groups work on this topic compared to the ones working on brain functions. Mainly neurophysiologists work on this subject and our originality is to combine physiology studies with mathematical modelling and experimental validation using our own neuroprostheses. The same analysis could be drawn with sensory feedback modelling. In this domain, our work is based on the recording and analysis of nerve activity through electro-neurography (ENG). We are interested in interpreting ENG in terms of muscle state in order to feedback useful information for FES controllers and to evaluate the stimulation effect. We believe that this knowledge should help to improve the design and programming of neuroprostheses. We investigate risky but promising fields such as intrafascicular recordings, area on which only few teams in North America (Canada and USA), and Denmark really work on. Very few teams in France, and none at Inria work on the peripheral nervous system modelling, together with experimental protocols that need neuroprostheses. Most of our Inria collaborators work on the central nervous system, except the spinal cord, (ODYSSEE for instance), or other biological functions (SISYPHE for instance). Our contributions concern the following aspects:

- Muscle modelling,
- Sensory organ modelling,
- Electrode nerve interface,
- High level motor function modelling,
- Model parameters identification.

We contribute both to the design of reliable and accurate experiments with a well-controlled environment, to the fitting and implementation of efficient computational methods derived for instance from Sigma Point Kalman Filtering.

3.2. Synthesis and Control of Human Functions

Participants: Christine Azevedo Coste, Philippe Fraisse, Mitsuhiro Hayashibe, David Andreu.

We aim at developping realistic solutions for real clinical problems expressed by patients and medical staff. Different approaches and specifications are developped to answer those issues in short, mid or long terms. This research axis is therefore obviously strongly related to clinical application objectives. Even though applications can appear very different, the problematic and constraints are usually similar in the context of electrical stimulation: classical desired trajectory tracking is not possible, robustness to disturbances is critical, possible observations of system are limited. Furthermore there is an interaction between body segments under voluntary control of the patient and body segments under artificial control. Finally, this axis relies on modelling and identification results obtained in the first axis and on the technological solutions and approaches developped in the third axis (Neuroprostheses). The robotics framework involved in DEMAR work is close to the tools used and developped by BIPOP team in the context of bipedal robotics. There is no national team working on those aspects. Within international community, several colleagues carry out researches on the synthesis and control of human functions, most of them belong to the International Functional Electrical Stimulation Society (IFESS) community. In the following we present two sub-objectives. Concerning spinal cord injuries (SCI) context not so many team are now involved in such researches around the world. Our force is to have technological solutions adapted to our theoretical developments. Concerning post-stroke context, several teams in Europe and North America are involved in drop-foot correction using FES. Our team specificity is to have access to the different expertises needed to develop new theoretical and technical solutions: medical expertise, experimental facilities, automatic control expertise, technological developments, industrial partner. These expertises are available in the team and through strong external collaborations.

3.3. Neuroprostheses

Participants: David Andreu, David Guiraud, Daniel Simon, Guy Cathébras, Fabien Soulier.

The main drawbacks of existing implanted FES systems are well known and include insufficient reliability, the complexity of the surgery, limited stimulation selectivity and efficiency, the non-physiological recruitment of motor units and muscle control. In order to develop viable implanted neuroprostheses as palliative solutions for motor control disabilities, the third axis "Neuroprostheses" of our project-team aims at tackling four main challenges: (i) a more physiologically based approach to muscle activation and control, (ii) a fibres' type and localization selective technique and associated technology (iii) a neural prosthesis allowing to make use of automatic control theory and consequently real-time control of stimulation parameters, and (iv) small, reliable, safe and easy-to-implant devices.

Accurate neural stimulation supposes the ability to discriminate fibres' type and localization in nerve and propagation pathway; we thus jointly considered multipolar electrode geometry, complex stimulation profile generation and neuroprosthesis architecture. To face stimulation selectivity issues, the analog output stage of our stimulus generator responds to the following specifications: i) temporal controllability in order to generate current shapes allowing fibres' type and propagation pathway selectivity, ii) spatial controllability of the current applied through multipolar cuff electrodes for fibres' recruitment purposes. We have therefore proposed and patented an original architecture of output current splitter between active poles of a multipolar electrode. The output stage also includes a monotonic DAC (Digital to Analog Converter) by design. However, multipolar electrodes lead to an increasing number of wires between the stimulus generator and the electrode contacts (poles); several research laboratories have proposed complex and selective stimulation strategies involving multipolar electrodes, but they cannot be implanted if we consider multisite stimulation (i.e. stimulating on several nerves to perform a human function as a standing for instance). In contrast, all the solutions tested on humans have been based on centralized implants from which the wires output to only monopolar or bipolar electrodes, since multipolar ones induce to many wires. The only solution is to consider a distributed FES architecture based on communicating controllable implants. Two projects can be cited: Bion technology (main competitor to date), where bipolar stimulation is provided by injectable autonomous units, and the LARSI project, which aimed at multipolar stimulation localized to the sacral roots. In both cases, there was no application breakthrough for reliable standing or walking for paraplegics. The power source, square stimulation shape and bipolar electrode limited the Bion technology, whereas the insufficient selection accuracy of the LARSI implant disqualified it from reliable use.

Keeping the electronics close to the electrode appears to be a good, if not the unique, solution for a complex FES system; this is the concept according to which we direct our neuroprosthesis design and development, in close relationship with other objectives of our project-team (control for instance) but also in close collaboration with medical and industrial partners.
Our efforts are mainly directed to implanted FES systems but we also work on surface FES architecture and stimulator; most of our concepts and advancements in implantable neuroprostheses are applicable somehow to external devices.

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:

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

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

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

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,

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

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:

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

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:

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

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:

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

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,

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

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 $f_{p_1p_2}(x_{p_1}, x_{p_2})$. 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:

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

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 $\{C_i, i \in [1, n] = \{\{p_{i^1}, \cdots, p_{i|C_i|}\}\}$ of order $|C_i|$ that will augment the definition of $\mathcal{V}$ and will introduce hyper-vertices:

$$\arg\min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}) + \sum_{C_i \in \mathcal{E}} f_{p_1 \cdots p_n}(x_{p_{i^1}}, \cdots, p_{x_{i|C_i|}}). \tag{12}$$

where $f_{p_1 \cdots p_n}$ is the price to pay for associating the labels $(x_{p_{i^1}}, \cdots, p_{x_{i|C_i|}})$ to the nodes $(p_1 \cdots p_{i|C_i|})$. 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.

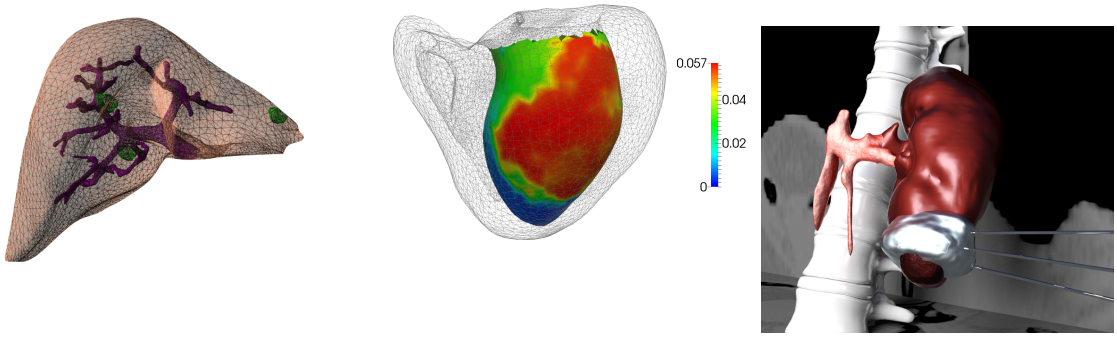
MIMESIS Team

3. Research Program

3.1. Real-Time Patient-Specific Computational Models

The main objective of this scientific challenge is the modeling of the biomechanics and physiology of certain organs under various stimuli. This requires describing different biophysical phenomena such as soft-tissue deformation, fluid dynamics, electrical propagation, or heat transfer. These models help simulate the impact of certain therapies (such as cryosurgery or radio-frequency ablation), but also represent the behavior of complex organs such as the brain, the liver or the heart. A common requirement across these developments is the need for (near) real-time computation and the ability to take into account for patient-specific characteristics.

An important part of our research was dedicated to the development of new **accurate models that remain compatible with real-time computation**. 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. Yet, patient-specific planning or per-operative guidance also requires the models to be **parametrized with patient-specific biomechanical data**. The objective in this area is related to the study of hyper-elastic models and their validation for a range of tissues. Preliminary work in this area has been done through two collaborations, one with the biomechanical lab in Lille (LML) with which we have a joint PhD student, and the biomechanics group from the ICube laboratory in Strasbourg on the development and validation of liver and kidney models. Another important research topic was related to model reduction through various approaches, such as Proper Generalized Decomposition (PGD). We have already established discussions with the LEGATO team at University of Luxembourg which has very good expertise in this area.



(a) Patient-specific mechanical model of the liver with its vascular system extracted from Computed Tomography Angiogram, and compatible with real-time simulation

(b) Simulation of the electrical activity of a human heart from intra-operative depolarization map

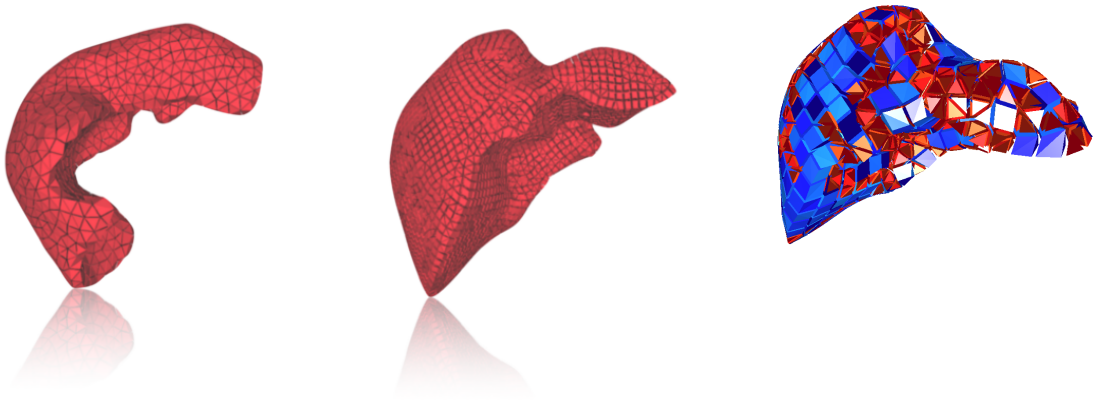
(c) Thermal evolution within a human kidney used in the context of cryoablation

Figure 3. Multi-physics and patient-specific models

We continued our work on cardiac electro-physiology simulation, with a focus on patient-specific adaptation of the model. We also studied the simulation of heat transfer and optimization problems in the context of heat diffusion. This work is a key element of the development of a planning system, such as for cryoablation procedures (Figure 3).

3.2. Adaptive Meshing and Advanced Simulation Technologies

The principal objective of this second challenge is to improve, at the numerical level, the efficiency, robustness, and quality of the simulations. To reach these goals, we followed two main directions: **adaptive meshing** to allow mesh transformations during a simulation and support cuts, local remeshing or dynamic refinement into areas of interest; and **numerical techniques**, such as asynchronous solvers, domain decomposition and model order reduction. 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 also considered constant. 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**. This covers various topics, such as simulation for cutting, tearing, fracture but also the use of multi-resolution meshes where elements are subdivided into areas where numerical errors need to be kept small [37], [38]. We also continued our 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 (Figure 4). This research also includes the study of domain decomposition techniques and other coupling techniques for multi-domain multi-physics simulations.



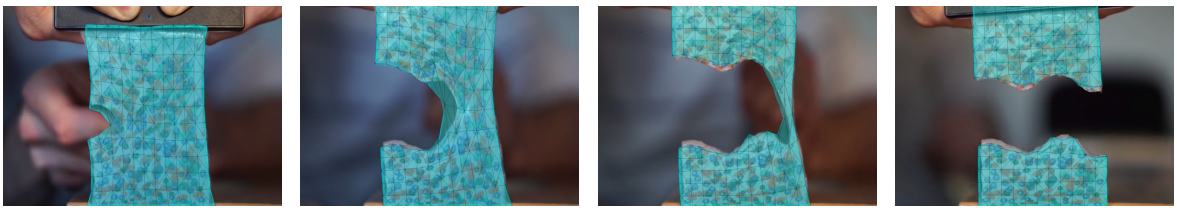
(a) Tetrahedral mesh (b) Hexahedral mesh (c) Mixed tetrahedral + hexahedral elements

Figure 4. Patient-specific volumetric meshes of the liver

3.3. Image-Driven Simulation

Image-guided 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 patient per-operative data 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. 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. 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 5) and led to several key publications [35], [36], [34] and awards (best paper at ISMAR 2013, second best paper at IPCAI 2014). We continued this work with an emphasis on robustness to uncertainty and outliers in the information extracted in real-time from image data. We also improved our 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.



(a) Step 1 (b) Step 2 (c) Step 3 (d) Step 4

Figure 5. An augmented elastic object undergoing large deformations and topological changes. The computation of the physics-based deformation, the cut detection and the topological modification of the underlying volumetric model are all performed in real-time.

The use of simulation in the context of image-guided therapy can be extended in several other ways. A direction we particularly want to address is the c**ombined 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), 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. The reduced-order stochastic filtering is a computationally efficient improvement over traditional computationally expensive **approaches which fits well the real-time and patient-specific requirements** arising from our per-operative context.

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 [61]. 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 [57], 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 undoubtly 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 [63], [50]. 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 heterogenous 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 mimick 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 homonculus or any central clock. Numerical schemes developed for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [47], [56] and were the basis for several contributions in our group [62], [59], [64]. 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 [48]; the associated formalism [53] 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 [45], that demonstrated a richness of behavior [49] 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 [59], [60] and their use for observing the emergence of typical cortical properties [52]. 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 (*cf.* § 6.1). 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, as we will discuss in § 6.1).

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 [55] 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 [51], [44]. 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 episods and supervising the training of a procedural memory in the cortex.

At the behavioral level, as mentionned 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 continous re-arrangement is supposed to involve several kinds of learning, at different time scales (from msec to years in humans) and to interfer 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 protocoles 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* [46]). 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 mimick 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.

NEUROMATHCOMP Project-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 occurring in the network and which phenomenological parameters control these bifurcations. Another issue concerns the interplay between neuron dynamics and synaptic network structure.

In this spirit, our team has been able to characterize the generic dynamics exhibited by models such as Integrate and Fire models [10], conductance-based Integrate and Fire models [58], [62], [5], models of epilepsy [92], effects of synaptic plasticity [88], [89], homeostasis and intrinsic plasticity [8].
Selected publications on this topic: lien.

3.2. Mean-field 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 50μm 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 [7] and the study of interacting diffusions [1]. Our investigations have shown that the rigorous dynamics mean-field equations can have a quite more complex structure than the ones commonly used in the literature (e.g. [74]) as soon as realistic effects such as synaptic variability are taken into account. Our goal is to relate those theoretical results with experimental measurement, especially in the field of optical imaging. For this we are collaborating with Institut des Neurosciences de la Timone, Marseille.
Selected publications on this topic.

3.3. Neural fields

Neural fields are a phenomenological way of describing the activity of population of neurons by delay 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 of noise.
Our group is developing mathematical and numerical methods for analysing these equations. These methods are based upon techniques from mathematical functional analysis [6], bifurcation theory [11], 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 [3], 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 [94]. This theory is based on center manifold and normal forms ideas. We are currently extending the theory to take into account various sources of noise using tools from the theory of stochastic partial differential equations.

Selected publications on this topic: lien.

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 [76] — and numerical methods such as pseudo-arclength continuation [67]. 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) [65], [75])*, which represent an alternation between subthreshold and spiking behaviour, and *bursting oscillations* [66], [73], also corresponding to experimentally observed behaviour [64].

Selected publications on this topic: lien.

3.5. Spike train statistics

The neuronal activity is manifested by the emission of action potentials (“spikes”) constituting spike trains. Those spike trains are usually not exactly reproducible when repeating the same experiment, even with a very good control ensuring that experimental conditions have not changed. Therefore, researchers are seeking models for spike train statistics, assumed to be characterized by a canonical probabilities giving the statistics of spatio-temporal spike patterns. A current goal in experimental analysis of spike trains is to approximate this probability from data. Several approach exist either based on (i) generic principles (maximum likelihood, maximum entropy); (ii) phenomenological models (Linear-Non linear, Generalized Linear Model, mean-field); (iii) Analytical results on spike train statistics in Neural Network models.

Our group is working on those 3 aspects, on a fundamental and on a practical (numerical) level. On the one hand, we have published analytical (and rigorous) results on statistics of spike trains in canonical neural network models (Integrate and Fire, conductance based with chemical and electric synapses) [2], [59], [5]. The main result is the characterization of spike train statistics by a Gibbs distribution whose potential can be explicitly computed using some approximations. Note that this result does not require an assumption of stationarity. We have also shown that the distributions considered in the cases (i), (ii), (iii) above are all Gibbs distributions [60]. On the other hand, we are proposing new algorithms for data processing . We have developed a C++ software for spike train statistics based on Gibbs distributions analysis and freely available at https://enas.inria.fr/. We are using this software in collaboration with several biologist groups involved in the analysis of retina spike trains (Centro de Neurociencia Valparaiso; Institut de la vision, Paris; Faculty of Medical Sciences, Newcastle University, Institute for Adaptive and Neural Computation, University of Edinburgh).
Selected publications on this topic: lien.

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 (Bienenstock, Cooper, and Munroe, 1982). 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 (convergence to a specific Gibbs distribution via a variational principle), resulting in a response-adaptation of the network to learned stimuli [88], [89], [61]. We are also studying the conjugated effects of synaptic and intrinsic plasticity in collaboration with H. Berry (Inria Beagle) and B. Delord, J. Naudé, ISIR team, Paris. On the other hand, we have pursued a geometric approach in which we show how a Hopfield network represented by a neural field with modifiable recurrent connections undergoing slow Hebbian learning can extract the underlying geometry of an input space [71]. We have also pursued an approach based on the ideas developed in the theory of slow-fast systems (in this case a set of neural fields equations) in the presence of noise and applied temporal averaging methods to recurrent networks of noisy neurons undergoing a slow and unsupervised modification of their connectivity matrix called learning [72].

Selected publications on this topic: lien.

3.7. Visual neuroscience

Our group focuses on the visual system to understand how information is encoded and processed resulting in visual percepts. To do so, we propose functional models of the visual system using a variety of mathematical formalisms, depending on the scale at which models are built, such as spiking neural networks or neural fields. So far, our efforts have been focused on the study of retinal processing, edge and texture perception, motion integration at the level of V1 and MT cortical areas.

At the retina level, we are modeling its circuitry [14] and we are studying the statistics of the spike train output (see, e.g., the software ENAS https://enas.inria.fr/). Real cell recordings are also analysed in collaboration with Institut de la vision, Paris; Centro de Neurociencia Valparaiso; Faculty of Medical Sciences, Newcastle University. For visual edges perception, we have used the theory of neural fields [13]. For visual textures perception, we have used a combination of neural fields theory and equivariant bifurcations theory [3]. At the level of V1-MT cortical areas, we have been investigating the temporal dynamics of motion integration for a wide range of visual stimuli [85], [90], [57], [9]. This work is done in collaboration with Institut des Neurosciences de la Timone, Marseille.

Selected publications on this topic: lien.

3.8. Neuromorphic vision

From the simplest vision architectures in insects to the extremely complex cortical hierarchy in primates, it is fascinating to observe how biology has found efficient solutions to solve vision problems. Pioneers in computer vision had this dream to build machines that could match and perhaps outperform human vision. This goal has not been reached, at least not on the scale that was originally planned, but the field of computer vision has met many other challenges from an unexpected variety of applications and fostered entirely new scientific and technological areas such as computer graphics and medical image analysis. However, modelling and emulating with computers biological vision largely remains an open challenge while there are still many outstanding issues in computer vision.

Our group is working on neuromorphic vision by proposing bio-inspired methods following our progress in visual neuroscience. Our goal is to bridge the gap between biological and computer vision, by applying our visual neuroscience models to challenging problems from computer vision such as optical flow estimation [91], coding/decoding approaches [80], [81] or classification [68], [69].

Selected publications on this topic: lien.

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.
Once having extracted models for different description levels, typically they are applied to obtain simulated activity which is supposed to reconstruct features in experimental data. Although this approach appears straight-forward, it implies 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 in a 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 monitor. This construction involves additional challenges and stipulates the contact to 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 to compare theoretical neural population activity 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 behaviour 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, the neuroimaging *inverse problem* consists in predicting individual information (behavior, phenotype) from neuroimaging data, while the *forward problem* consists in fitting neuroimaging data with high-dimensional (e.g. genetic) 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 reasonably well ?), and a regularization schemes that represents a prior on the expected solution of the problem. In particular some priors enforce some properties of the solutions, such as sparsity, smoothness or being piece-wise constant.
Let us detail the model used in the inverse problem: Let $\mathbf{X}$ be a neuroimaging dataset as an (n_{subj}, n_{voxels}) matrix, where n_{subj} and n_{voxels} are the number of subjects under study, and the image size respectively, $\mathbf{Y}$ an array of values that represent characteristics of interest in the observed population, written as (n_{subj}, n_f) matrix, where n_f is the number of characteristics that are tested, and β an array of shape (n_{voxels}, n_f) that represents a set of pattern-specific maps. In the first place, we may consider the columns $\mathbf{Y}_1, .., \mathbf{Y}_{n_f}$ of Y independently, yielding n_f problems to be solved in parallel:

$$\mathbf{Y}_i = \mathbf{X}\beta_i + \epsilon_i, \forall i \in \{1, .., n_f\},$$

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

$$\widehat{\beta_i} = \text{argmin}_{\beta_i} \|\mathbf{Y}_i - \mathbf{X}\beta_i\|^2 + \Psi(\beta_i), \tag{13}$$

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

$$\Psi(\beta; \lambda_1, \lambda_2, \eta_1, \eta_2) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2 + \eta_1 \|\nabla\beta\|_1 + \eta_2 \|\nabla\beta\|_2 \tag{14}$$

with λ_1, λ_2, η_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 β 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 ($\eta_2 > 0$ and $\lambda_1 > 0$ only), and yields smooth, compactly supported spatial basis functions.

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 $i \in \{1, .., n_f\}$. This can be computed through cross-validation, by *learning* $\widehat{\beta_i}$ on some part of the dataset, and then estimating $(Y_i - X\widehat{\beta_i})$ using the remainder of the dataset.

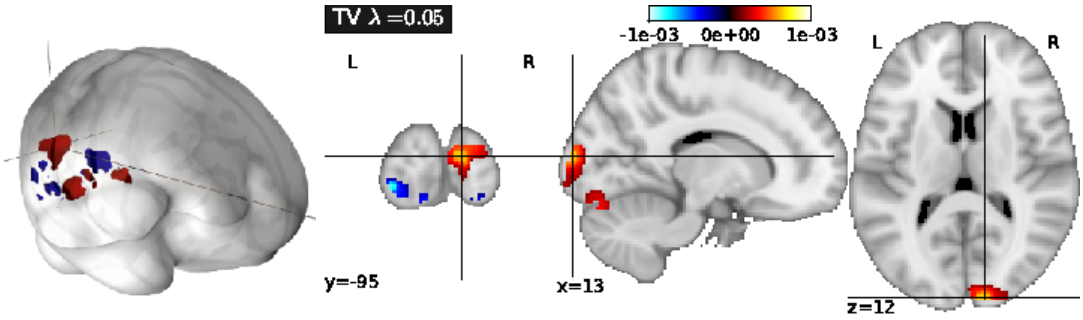


Figure 1. Example of the regularization of a brain map with total variation in an inverse problem. The problem here consists in predicting the spatial scale of an object presented as a stimulus, given functional neuroimaging data acquired during the observation of an image. Learning and test are performed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

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 is particularly important to include external anatomical priors on the relevant 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 regression*, where a logistic 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* is an important question, as it makes little sense that a learned model depends dramatically on the particular observations used for learning. This is an important issue, as this kind of robustness is somewhat opposite to sparsity requirements.
- *Multi-task learning*: if several target variables are thought to be related, it might be useful to constrain the estimated parameter vector β 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 β_i and thus, improve the prediction of $\mathbf{Y}_i$. Yet this does not impose constraints on the non-zero parameters of the parameters β_i.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{15}$$

then

$$\widehat{\beta} = \text{argmin}_{\beta=(\beta_i),i=1..n_f} \sum_{i=1}^{n_f} \|\mathbf{Y_i} - \mathbf{X}\beta_\mathbf{i}\|^2 + \lambda \sum_{j=1}^{n_{voxels}} \sqrt{\sum_{i=1}^{n_f} \beta_{\mathbf{i,j}}^{\mathbf{2}}} \tag{16}$$

3.2. Multivariate decompositions

Multivariate decompositions are an important tool 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 depicting similar activities 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_{subj}, n_{voxels}) matrix, after proper centering; the model reads

$$\mathbf{X} = \mathbf{AD} + \epsilon, \tag{17}$$

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 reconstruct 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:

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

where $(\mathbf{A}_i),\ i \in \{1..n_f\}$ 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).

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 assess whether an observation is aberrant or not or in classification problems. 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 an important question that needs to be addressed.
- **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-samples 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.

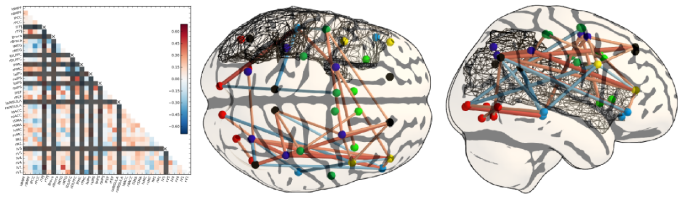


Figure 2. Example of functional connectivity analysis: The correlation matrix describing brain functional connectivity in a post-stroke patient (lesion volume outlined as a mesh) is compared to a group of control subjects. Some edges of the graphical model show a significant difference, but the statistical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

POPIX Team

3. Research Program

3.1. Research Program

Mathematical models that characterize complex biological phenomena are complex numerical models which are defined by systems of ordinary differential equations when dealing with dynamical systems that evolve with respect to time, or by partial differential equations when there is a spatial component to the model. Also, it is sometimes useful to integrate a stochastic aspect into the dynamical systems in order to model stochastic intra-individual variability.

In order to use such methods, we are rapidly confronted with complex numerical difficulties, generally related to resolving the systems of differential equations. Furthermore, to be able to check the quality of a model, we require data. The statistical aspect of the model is thus critical in its way of taking into account different sources of variability and uncertainty, especially when data comes from several individuals and we are interested in characterizing the inter-subject variability. Here, the tool of reference is mixed-effects models.

Mixed-effects models are statistical models with both fixed effects and random effects, i.e., mixed effects. They are useful in many real-world situations, especially in the physical, biological and social sciences. In particular, they are well-adapted to situations where repeated measurements are made on the same individual/statistical unit.

POPIX develops new methods for estimation of complex mixed-effects models. Some of the extensions to these models that POPIX is actively researching include:

- models defined by a large system of differential equations
- models defined by a system of stochastic differential equations
- models defined by partial differential equations
- mixed hidden Markov models
- mixture models and model mixtures
- time-to-event models
- models including a large number of covariates

It is also important to clarify that POPIX is not meant to be a team of modelers; our main activity is not to develop models, but to develop tools for modelers. Indeed, we are of course led via our various collaborations to interact closely with modelers involved in model development, in particular in the case of our collaborations with modeling and simulation teams in the pharmaceutical industry. But POPIX is not in the business of building PKPD models per se.

Lastly, though pharmacometrics remains the main field of interest for the population approach, this approach is also appropriate to address other types of complex biological phenomena exhibiting inter-individual variability and necessitating therefore to be described by numerical and statistical models. We have already demonstrated the relevance of the developed approaches and tools in diverse other domains such as agronomy for characterizing corn production, and cellular biology for characterizing the cell cycle and the creation of free radicals in cells. Now we wish to push on to explore new areas of modeling such as for the respiratory system and blood flow. But again, it is not within the scope of the activities of POPIX to develop new models; instead, the goal is to demonstrate the relevance of the population approach in these areas.

SISTM Project-Team

3. Research Program

3.1. Mecanistic 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 summarizing the dynamics over time in term of slopes of the trajectories [37]. These slopes can be compared between treatment groups or according to patients' characteristics. Another way for analyzing 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 [30]. 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 [29].

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 [22]. 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 [21].

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 [33], [40], [25], [31], [26], [39], 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 became 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 became 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 network [0]. As an exemple, 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 [36]. 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.

[0] see 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 [34]. 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 [38]. 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 [27]. 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...

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 endorse 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 the Neurobase national project where we had a leading role, we wanted to enhance the development of distributed and heterogeneous medical image processing systems.

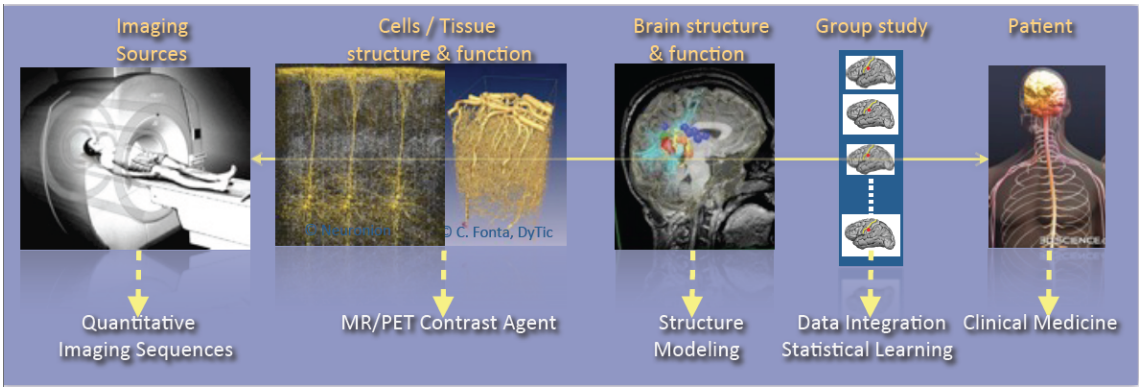


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales : from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions,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

As shown in figure 1 , research activities of the VISAGES U746 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 includes:

- 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 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 7.1.1) 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 7.2.2 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). 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 sense of the models (uniqueness, conservativity, entropy dissipation, ...) and propose effective numerical methods to estimate their solution in complex configurations (multi-dimensional, unstructured meshes, well-balanced, ...).

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 discretization 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, 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 must rely on a special analysis.

3.2.1. Multilayer approach

As for the first shallow water assumption, *multi-layer* systems were proposed describing the flow as a superposition of Saint-Venant type systems [39], [43], [45]. Even if this approach has provided interesting results, layers are considered separate and non-miscible fluids, which imply strong limitation. That is why we proposed a slightly different approach [40], [41] based on Galerkin type decomposition along the vertical axis of all variables and leading, both for the model and its discretization, to more accurate results.

A kinetic representation of our multilayer model allows to derive robust numerical schemes endowed with 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 enable to guess the optimal number of layers.
- The introduction of several source terms due for instance to Coriolis forces 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 proved and conversely the possible loss of hyperbolicity cannot be characterized. 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.

When 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 integrating the 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-physic 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 zone 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 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 system Euler/Poisson, 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 such as threshold laws.
- Interaction between flows and floating structures is the challenge at the scale of the shallow water equations. This study needs 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 scheme. 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.

3.3. 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 has to 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

The hydrodynamic models comprise advection and sources terms. The conservation of the balance between the source terms, typically viscosity and friction, has a significant impact since the overall flow is generally a perturbation around one 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 potential waves and the physical velocity makes classical numerical schemes not efficient: 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 potential 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. Data assimilation

Data assimilation consists in a coupling between a model and observation measurements. Developing robust data assimilation methods for hyperbolic-type conservation laws is a challenging subject. These PDEs indeed show no dissipation effects and the input of additional information in the model equations may introduce errors that propagate and create shocks. We have recently proposed a new approach based on the kinetic description of the conservation law. Hence, data assimilation is carried out at the kinetic level, using a Luenberger observer. Assimilation then resumes to the handling of a BGK type equation. The advantage of this framework is that we deal with a single "linear" equation instead of a nonlinear system and it is easy to recover the macroscopic variables. We are able to prove the convergence of the model towards the data in case of complete observations in space and time.

This work is done in collaboration with the M3DISIM Inria project-team. M. Doumic and B. Perthame (MAMBA) also participate.

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 reseach topics are:

1. Modelling and analysis
   - Fluid closure in plasma
   - Turbulence
   - Plasma anisotropy type instabilities
   - Coupling FBE – Transport
   - Halo current
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
   - Ionospheric plasma

CLIME Project-Team

3. Research Program

3.1. Data assimilation and inverse modeling

This activity is one major concern of environmental sciences. It matches up the setting and the use of data assimilation methods, for instance variational methods (such as the 4D-Var method). An emerging issue lies in the propagation of uncertainties by models, notably through ensemble forecasting methods.

Although modeling is not part of the scientific objectives of Clime, the project-team has complete access to air quality models through collaborations with École des Ponts ParisTech and EDF R&D: the models from Polyphemus (pollution forecasting from local to regional scales) and Code_Saturne (urban scale). In regard to other modeling domains, such as oceanography and meteorology, Clime accesses models through co-operation with LOCEAN (Laboratoire d'OCEANographie et du climat, UPMC) and Météo-France.

The research activities of Clime tackle scientific issues such as:

- Within a family of models (differing by their physical formulations and numerical approximations), which is the optimal model for a given set of observations?
- How to reduce dimensionality of problems by Galerkin projection of equations on subspaces? How to define these subspaces in order to keep the main properties of systems?
- How to assess the quality of a forecast and its uncertainty? How do data quality, missing data, data obtained from sub-optimal locations, affect the forecast? How to better include information on uncertainties (of data, of models) within the data assimilation system?
- How to make a forecast (and a better forecast!) by using several models corresponding to different physical formulations? It also raises the question: how should data be assimilated in this context?
- Which observational network should be set up to perform a better forecast, while taking into account additional criteria such as observation cost? What are the optimal location, type and mode of deployment of sensors? How should trajectories of mobile sensors be operated, while the studied phenomenon is evolving in time? This issue is usually referred as "network design".

3.2. Satellite acquisitions and image assimilation

In geosciences, the issue of coupling data, in particular satellite acquisitions, and models is extensively studied for meteorology, oceanography, chemistry-transport and land surface models. However, satellite images are mostly assimilated on a point-wise basis. Three major approaches arise if taking into account the spatial structures, whose displacement is visualized on image sequences:

- Image approach. Image assimilation allows the extraction of features from image sequences, for instance motion field or structures' trajectory. A model of the dynamics is considered (obtained by simplification of a geophysical model such as Navier-Stokes equations). An observation operator is defined to express the links between the model state and the pixel values or some image features. In the simplest case, the pixel value corresponds to one coordinate of the model state and the observation operator is reduced to a projection. However, in most cases, this operator is highly complex, implicit and non-linear. Data assimilation techniques are developed to control the initial state or the whole assimilation window. Image assimilation is also applied to learn reduced models from image data and estimate a reliable and small-size reconstruction of the dynamics, which is observed on the sequence.
- Model approach. Image assimilation is used to control an environmental model and obtain improved forecasts. In order to take into account the spatial and temporal coherency of structures, specific image characteristics are considered and dedicated norms and observation error covariances are defined.

- Correcting a model. Another topic, mainly described for meteorology in the literature, concerns the location of structures. How to force the existence and to correct the location of structures in the model state using image information? Most of the operational meteorological forecasting institutes, such as Météo-France (in France), UK-met (in United Kingdom), KNMI (in Netherlands), ZAMG (in Austria) and Met-No (in Norway), study this issue because operational forecasters often modify their forecasts based on visual comparisons between the model outputs and the structures displayed on satellite images.

3.3. Software chains for environmental applications

An objective of Clime is to participate in the design and creation of software chains for impact assessment and environmental crisis management. Such software chains bring together static or dynamic databases, data assimilation systems, forecast models, processing methods for environmental data and images, complex visualization tools, scientific workflows, ...

Clime is currently building, in partnership with École des Ponts ParisTech and EDF R&D, such a system for air pollution modeling: Polyphemus (see the web site http://cerea.enpc.fr/polyphemus/), whose architecture is specified to satisfy data requirements (e.g., various raw data natures and sources, data preprocessing) and to support different uses of an air quality model (e.g., forecasting, data assimilation, ensemble runs).

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 arises 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 [25] 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 [38]. 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, flurocein,...). 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 including as constraints 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.

A substantial progress has been done in this direction with the design of dedicated dense estimation techniques to estimate dense fluid motion fields[4], [11], the setting up of tomographic techniques to carry out 3D velocity measurements [32], the inclusion of physical constraints to infer 3D motions or the design of dynamically consistent velocity measurements to provide coherent motion fields from time resolved fluid flow image sequences [10]. These progresses have brought further accuracy and an improved spatial resolution for a variety of applications ranging from experimental fluid mechanics to geophysical sciences. For a detailed review of these approaches see [7].

We believe that such approaches must be first enlarged to the wide variety of imaging modalities enabling the observation of fluid flows. This covers for instance, the systematic study of motion estimation for the different channels of meteorological satellites, but also of other experimental imaging tools such as Shadowgraphs, Background oriented Schlieren, Schlieren [45], diffusive scalar images, fluid holography [46], or Laser Induced Fluorimetry. All these modalities offer the possibility to visualize time resolved sequences of the flow. The velocity measurement processes available to date for that kind of images suffer from a lack of physical relevancy to keep up with the increasing amount of fine and coherent information provided by the images. We think, and have begun to prove, that a significant step forward can be taken by providing new tools based on sound data models and adapted regularization functional, both built on physical grounds.

Additional difficulties arise when considering the necessity to go towards 3D measurements and 3D volumetric reconstruction of the observed flows (e.g., the tomographic PIV paradigm). First, unlike in the standard setup, the 2D images captured by the experimentalists only provide a partial information about the structure of the particles transported by the fluid. As a matter of fact, inverse problems have to be solved in order to recover this crucial information. Secondly, another issue stands in the increase of the underdetermination of the problem, that is the important decrease of the ratio between the number of observations and the total number of unknowns. In particular, this point asks for methodologies able to gather and exploit observations

captured at different time instants. Finally, the dimensions of the problem (that is, the number of unknown) dramatically increase with the transition from the 2D to the 3D paradigm. This leads, as a by-product, to a significant amplification of the computational burden and requires the conception of efficient algorithms, exhibiting a reasonable scaling with the problem dimensions.

The first problem can be addressed by resorting to state-of-the-art methodologies pertaining to sparse representations. These techniques consist in identifying the solution of an inverse problem with the most “zero" components which, in the case of the tomographic PIV, turns out to be a physically relevant option. Hence, the design of sparse representation algorithms and the study of their conditions of success constitute an important research topic of the group. On the other hand, we believe that the dramatic increase of the under-determination appearing in the 3D setup can be tackled by combining tomographic reconstruction of several planar views of the flow with data assimilation techniques. These techniques enable to couple a dynamical model with incomplete observations of the flow. Each applicative situation under concern defines its proper required scale of measurement and a scale for the dynamical model. For instance, for control or monitoring purposes, very rapid techniques are needed whereas for analysis purpose the priority is to get accurate measurements of the smallest motion scales as possible. These two extreme cases imply the use of different models but also of different algorithmic techniques. Recursive techniques and large scale representation of the flow are relevant for the first case whereas batch techniques relying on the whole set of data available and models refined down to small scales have to be used for the latter case.

The question of the scale of the velocity measurement is also an open question that must be studied carefully. Actually, no scale considerations are taken into account in the estimation schemes. It is more or less abusively assumed that the measurements supplied have a subpixel accuracy, which is obviously erroneous due to implicit smoothness assumptions made either in correlation techniques or in variational estimation techniques. We are convinced that to go towards the measurement of the smaller scales of the flow it is necessary to introduce some turbulence or uncertainty subgrid modeling within the estimation scheme and also to devise alternative regularization schemes that fit well with phenomenological statistical descriptions of turbulence described by the velocity increments moments. As a by product such schemes should offer the possibility to have a direct characterization, from image sequences, of the flow turbulent regions in term of vortex tube, area of pure straining, or vortex sheet. This philosophy should allow us to elaborate methods enabling the estimation of relevant characteristics of the turbulence like second-order structure functions, mean energy dissipation rate, turbulent viscosity coefficient, or dissipative scales.

We are planning to study these questions for a wide variety of application domains ranging from experimental fluid mechanics to geophysical sciences. We believe there are specific needs in different application domains that require clearly identified developments and modeling. Let us for instance mention meteorology and oceanography which both involve very specific dynamical modeling but also micro-fluidic applications or bio-fluid applications that are ruled by other types of dynamics.

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) [3], [5].

Conversely, the injection of experimental data into simulations ensures some fitting of the model with reality. When used with the correct level of expertise to calibrate the models at the relevant scales, regarding data validity and the targeted representation scale, this collaboration represents a powerful tool for the analysis and reconstruction of the flows. Automated back and forth sequencing between data integration and calculations have to be elaborated for the different types of flows with a correct adjustment of the observed and modeled scales. This appears more and more feasible when considering the sensitivity, the space resolution and above all the time resolution that the imaging sensors are reaching now.

That becomes particularly true, for instance, for satellite imaging, the foreseeable advances of which will soon give the right complement to the progresses in atmospheric and ocean modeling to dramatically improve the analysis and predictions of physical states and streams for weather and environment monitoring. In that domain, there is a particular interest in being able to combine image data, models and in-situ measurements, as high densities of data supplied by meteorological stations are available only for limited regions of the world, typically Europe and USA, while Africa, or the south hemisphere lack of refined and frequent *in situ* measurements. Moreover, we believe that such an approach can favor great advances in the analysis and prediction of complex flows interactions like those encountered in sea-atmosphere interactions, dispersion of polluting agents in seas and rivers, etc. In other domains we believe that image data and dynamical models coupling may bring interesting solutions for the analysis of complex phenomena which involve multi-phasic flows, interaction between fluid and structures, and the general case of flows with complex unknown border conditions.

The coupling approach can be extended outside the fluidics domain to complex dynamics that can be modeled either from physical laws or from learning strategies based on the observation of previous events [1]. This concerns for instance forest combustion, the analysis of the biosphere evolution, the observation and prediction of the melting of pack ice, the evolution of sea ice, the study of the consequences of human activity like deforestation, city growing, landscape and farming evolution, etc. All these phenomena are nowadays rapidly evolving due to global warming. The measurement of their evolution is a major societal interest for analysis purpose or risk monitoring and prevention.

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, the tools provided by the Bayesian theory play a crucial role since they encompass a variety of methodologies to model and process uncertainty. As a consequence, the Bayesian paradigm 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 [40], [41] or through stochastic filtering approaches [33], [36]. 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 [29] 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 [48], [35]. 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 [26]. Most of the works of the literature indeed comes back to open-loop control approaches [43], [37], [42] or to forcing approaches [34] 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.

Another important part of the works conducted in Fluminance concerns the study of closed-loop approaches, for which the convergence of the system to a target state is ensured even in the presence of errors (related either to the flow model, the actuators, or the sensors) [31]. However, designing a closed loop control law requires the use of sensors that are both non-intrusive, accurate and adapted to the time and spacial scales of the phenomenon to monitor. Such sensors are unfortunately hardly available in the context of flow control. The only sensors currently used are wall sensors located in a limited set of measurement points [27], [30]. The difficulty is then to reconstruct the entire state of the controlled system from a model based only on the few measurements available on the walls [39]. Instead of relying on sparse measurements, we propose to use denser features estimated from images. With the capabilities of up-to-date imaging sensors, we can expect an improved reconstruction of the flow (both in space and time) enabling the design of efficient image based control laws. This formulation is referred to as visual servoing control scheme.

Visual servoing is a widely used technique for robot control. It consists in using data provided by a vision sensor for controlling the motions of a robot [28]. This technique, historically embedded in the larger domain of sensor-based control [44], can be properly used to control complex robotic systems or, as we showed it recently, flows [47].

Classically, to achieve a visual servoing task, a set of visual features, $\mathbf{s}$, has to be selected from visual measurements, $\mathbf{m}$, extracted from a current image. A control law is then designed so that these visual features reach a desired value, $\mathbf{s}^*$, related to the target state of the system. The control principle consists in regulating to zero the error vector: $\mathbf{e} = \mathbf{s} - \mathbf{s}^*$. To build the control law, the knowledge of the so-called *interaction matrix* $\mathbf{L_s}$ is usually required. This matrix links the time variation of $\mathbf{s}$ to the signal command $\mathbf{u}$. However, computing this matrix in the context of flow control is far more complex than in the case of robot control as flows are associated to chaotic nonlinear systems living in infinite dimensional spaces. As such, it is possible to formalize the model through a Galerkin projection in terms of an ODE system for which classical control laws can be applied. It is also possible to express the system with finite difference approximations and to use discrete time control algorithms amenable to modern micro-controllers. Alternatively, one may develop control methods directly on the infinite dimensional system and then finally discretize the resulting process for implementation purpose. Each approach has its own advantages and drawbacks. For the first two, known control methods can be used at the expense of a great sensibility to space discretization. The last one is less sensitive to discretization errors but more difficult to set up. These practical issues and their related theoretical difficulties make this study a very interesting field of research.

LEMON Team

3. Research Program

3.1. State of the Art

3.1.1. Shallow Water Models

Shallow Water (SW) wave dynamics and dissipation represent an important research field. This is because shallow water flows are the most common flows in geophysics. In shallow water regions, dispersive effects (non-hydrostatic pressure effects related to strong curvature in the flow streamlines) can become significant and affect wave transformations. The shoaling of the wave (the "steepening" that happens before the breaking) cannot be described with the usual Saint-Venant equations. To model such various evolutions, one has to use more sophisticated models (Boussinesq, Green-Naghdi...). Nowadays, the classical Saint-Venant equations can be solved numerically in an accurate way, allowing the generation of bores and the shoreline motion to be handled, using recent finite-volume or discontinuous-Galerkin schemes. In contrast, very few advanced works regarding the derivation and modern numerical solution of dispersive equations [29], [33], [60] are available in one dimensions, let alone in the multidimensional case. We can refer to [59], [36] for some linear dispersive equations, treated with finite-element methods, or to [33] for the first use of advanced high-order compact finite-volume methods for the Serre equations. Recent work undertaken during the ANR MathOCEAN [29] lead to some new 1D fully nonlinear and weakly dispersive models (Green-Naghdi like models) that allow to accurately handle the nonlinear waves transformations. High order accuracy numerical methods (based on a second-order splitting strategy) have been developed and implemented, raising a new and promising 1D numerical model. However, there is still a lack of new development regarding the multidimensional case.
In shallow water regions, depending on the complex balance between non-linear effects, dispersive effects and energy dissipation due to wave breaking, wave fronts can evolve into a large range of bore types, from purely breaking to purely undular bore. Boussinesq or Green-Naghdi models can handle these phenomena [27] . However, these models neglect the wave overturning and the associated dissipation, and the dispersive terms are not justified in the vicinity of the singularity. Previous numerical studies concerning bore dynamics using depth-averaged models have been devoted to either purely broken bores using NSW models [30], or undular bores using Boussinesq-type models [40]. Let us also mention [38] for tsunami modeling and [37], [49] for the dam-break problem. A model able to reproduce the various bore shapes, as well as the transition from one type of bore to another, is required. A first step has been made with the one-dimensional code [29], [57]. The SWASH project led by Zijlema at Delft [60] addresses the same issues.

3.1.2. Open boundary conditions and coupling algorithms

For every model set in a bounded domain, there is a need to consider boundary conditions. When the boundaries correspond to a modeling choice rather than to a physical reality, the corresponding boundary conditions should not create spurious oscillations or other unphysical behaviour at the artificial boundary. Such conditions are called **open boundary conditions** (OBC). They have been widely studied by applied mathematicians since the pionneering work of [39] on transparent boundary conditions. Deep studies of these operators have been performed in the case of linear equations, [44], [28], [54]. Unfortunately, in the case of geophysical fluid dynamics, this theory leads to nonlocal conditions (even in linear cases) that are not usable in numerical models. Most of current models (including high quality operational ones) modestly use a *no flux* condition (namely an homogeneous Neumann boundary condition) when a free boundary condition is required. But in many cases, Neumann homogeneous conditions are a very poor approximation of the exact transparent conditions. Hence the need to build higher order approximations of these conditions that remain numerically tractable.

Numerous physical processes are involved in coastal modeling, each of them depending on others (surface winds for coastal oceanography, sea currents for sandbars dynamics, etc.). Connecting two (or more) model solutions at their interface is a difficult task, that is often addressed in a simplified way from the mathematical viewpoint: this can be viewed as the one and only iteration of an iterative process. This results with a low quality coupled system, which could be improved either with additional iterations, and/or thanks to the improvement of interface boundary conditions and the use of OBC (see above). Promising results have been obtained in the framework of **ocean-atmosphere coupling** (in a simplified modeling context) in [50], where the use of advanced coupling techniques (based on domain decomposition algorithm) are introduced.

3.1.3. A need for upscaled shallow water models.

The mathematical modeling of **fluid-biology** coupled systems in lagoon ecosystems requires one or several water models. It is of course not necessary (and not numerically feasible) to use accurate non-hydrostatic turbulent models to force the biological processes over very long periods of time. There is a compromise to be reached between accurate (but untractable) fluid models such as the Navier-Stokes equations and simple (but imprecise) models such as [41].
In urbanized coastal zones, upscaling is also a key issue. This stems not only from the multi-scale aspects dealt with in the previous subsection, but also from modeling efficiency considerations.

The typical size of the relevant hydraulic feature in an urban area is between 0.1 m and 1.0 m, while the size of an urban area usually ranges from 10^3 m to 10^4 m. Refined flow computations (e.g. in simulating the impact of a tsunami) over entire coastal conurbations using a 2D horizontal model thus require 10^6 to 10^9 elements. From an engineering perspective, this makes both the CPU and man-supervised mesh design efforts unaffordable in the present state of technology.

Upscaling provides an answer to this problem by allowing macroscopic equations to be derived from the small-scale governing equations. The powerful, multiple scale expansion-based homogeneization technique [26], [25], [53] has been applied successfully to flow and transport upscaling in porous media, but its use is subordinated to the stringent assumptions of (i) the existence of a Representative Elementary Volume (REV), (ii) the scale separation principle, and (iii) the process is not purely hyperbolic at the microscopic scale, otherwise precluding the study of transient solutions [26]. Unfortunately, the REV has been shown recently not to exist in urban areas [43]. Besides, the scale separation principle is violated in the case of sharp transients (such as tsunami waves) impacting urban areas because the typical wavelength is of the same order of magnitude as the microscopic detail (the street/block size). Moreover, 2D shallow water equations are essentially hyperbolic, thus violating the third assumption.

These hurdles are overcome by averaging approaches. Single porosity-based, macroscopic shallow water models have been proposed [35], [42], [45] and applied successfully to urban flood modeling scale experiments [42], [51], [56]. They allow the CPU time to be divided by 10 to 100 compared to classical 2D shallow water models. Recent extensions of these models have been proposed in the form of integral porosity [55] and multiple porosity [43] shallow water models.

3.2. Scientific Objectives

Our main challenge is: build and couple elementary models in coastal areas to improve their capacity to simulate complex dynamics. This challenge consists of three principal scientific objectives. First of all, each of the elementary models has to be consistently developed (regardless of boundary conditions and interactions with other processes). Then open boundary conditions (for the simulation of physical processes in bounded domains) and links between the models (interface conditions) have to be identified and formalized. Finally, models and boundary conditions (*i.e.* coupled systems) should be proposed, analyzed and implemented in a common platform.

3.2.1. Single process models and boundary conditions

The time-evolution of a water flow in a three-dimensional computational domain is classically modeled by Navier-Stokes equations for incompressible fluids. Depending on the physical description of the considered domain, these equations can be simplified or enriched. Consequently, there are **numerous water dynamics models** that are derived from the original Navier-Stokes equations, such as primitive equations, shallow water equations (see [34]), Boussinesq-type dispersive models [27]), etc. The aforementioned models have **very different mathematical natures**: hyperbolic *vs* parabolic, hydrostatic *vs* non-hydrostatic, inviscid *vs* viscous, etc. They all carry nonlinearities that make their mathematical study (existence, uniqueness and regularity of weak and/or strong solutions) highly challenging (not to speak about the $1M Clay competition for the 3D Navier Stokes equations, which may remain open for some time).

The objective is to focus on the mathematical and numerical modeling of models adapted to **nearshore dynamics**, accounting for complicated wave processes. There exists a large range of models, from the shallow water equations (eventually weakly dispersive) to some fully dispersive deeper models. All these models can be obtained from a suitable asymptotic analysis of the water wave equations (Zakharov formulation) and if the theoretical study of these equations has been recently investigated [48], there is still some serious numerical challenges. So we plan to focus on the derivation and implementation of robust and high order discretization methods for suitable two dimensional models, including enhanced fully nonlinear dispersive models and fully dispersive models, like the Matsuno-generalized approach proposed in [47]. Another objective is to study the shallow water dispersive models without any irrotational flow assumption. Such a study would be of great interest for the study of nearshore circulation (wave induced rip currents).

For obvious physical and/or computational reasons, our models are set in bounded domains. Two types of boundaries are considered: physical and mathematical. Physical boundaries are materialized by an existing interface (atmosphere/ocean, ocean/sand, shoreline, etc.) whereas mathematical boundaries appear with the truncation of the domain of interest. In the latter case, **open boundary conditions** are mandatory in order not to create spurious reflexions at the boundaries. Such boundary conditions being nonlocal and impossible to use in practice, we shall look for approximations. We shall obtain them thanks to the asymptotic analysis of the (pseudo-differential) boundary operators with respect to small parameters (viscosity, domain aspect ratio, Rossby number, etc.). Naturally, we **will seek the boundary conditions leading to the best compromise** between mathematical well-posedness and physical consistency. This will make extensive use of the mathematical theory of **absorbing operators** and their approximations [39].

3.2.2. Coupled systems

The Green-Naghdi equations provide a correct description of the waves up to the breaking point while the Saint-Venant equations are more suitable for the description of the surf zone (i.e. after the breaking). Therefore, the challenge here is first to **design a coupling strategy** between these two systems of equations, first in a simplified one-dimensional case, then to the two-dimensional case both on cartesian and unstructured grids. High order accuracy should be achieved through the use of flexible Discontinuous-Galerkin methods.
Additionally, we will couple our weakly dispersive shallow water models to other fully dispersive deeper water models. We plan to mathematically analyze the coupling between these models. In a first step, we have to understand well the mixed problem (initial and boundary conditions) for these systems. In a second step, these new mathematical development have to be embedded within a numerically efficient strong coupling approach. The deep water model should be fully dispersive (solved using spectral methods, for instance) and the shallow-water model will be, in a first approach, the Saint-Venant equations. Then, when the 2D extension of the currently developped Green-Naghdi numerical code will be available, the improved coupling with a weakly dispersive shallow water model should be considered.

In the context of Schwarz relaxation methods, usual techniques can be seen as the first iteration (not converged) of an iterative algorithm. Thanks to the work performed on efficient boundary conditions, we shall **improve the quality of current coupling algorithms**, allowing for qualitatively satisfying solutions **with a reduced computational cost** (small number of iterations).

We are also willing to explore the role of geophysical processes on some biological ones. For example, the design of optimal shellfish farms relies on confinement maps and plankton dynamics, which strongly depend on long-time averaged currents. Equations that model the time evolution of species in a coastal ecosystem are relatively simple from a modeling viewpoint: they mainly consist of ODEs, and possibly advection-diffusion equations. The issue we want to tackle is the choice of the fluid model that should be coupled to them, accounting for the important time scales discrepancy between biological (evolution) processes and coastal fluid dynamics. Discrimination criteria between refined models (such as turbulent Navier-Stokes) and cheap ones (see [41]) will be proposed.

Coastal processes evolve at very different time scales: atmosphere (seconds/minutes), ocean (hours), sediment (months/years) and species evolution (years/decades). Their coupling can be seen as a *slow-fast* dynamical system, and a naïve way to couple them would be to pick the smallest time-step and run the two models together: but the computational cost would then be way too large. Consequently **homogenization techniques or other upscaling methods** should be used in order to account for these various time scales at an affordable computational cost. The research objectives are the following:

- So far, the proposed upscaled models have been validated against theoretical results obtained from refined 2D shallow water models and/or very limited data sets from scale model experiments. The various approaches proposed in the literature [31], [32], [35], [42], [43], [45], [51], [55], [56] have not been compared over the same data sets. Part of the research effort will focus on the extensive validation of the models on the basis of scale model experiments. Active cooperation will be sought with a number of national and international Academic partners involved in urban hydraulics (UCL Louvain-la-Neuve, IMFS Strasbourg, Irvine University California) with operational experimental facilities.
- Upscaling of source terms. Two types of source terms play a key role in shallow water models: geometry-induced source terms (arising from the irregular bathymetry) and friction/turbulence-induced energy loss terms. In all the upscaled shallow water models presented so far, only the large scale effects of topographical variations have been upscaled. In the case of wetting/drying phenomena and small depths (e.g. the *Camargue* tidal flats), however, it is forseen that subgrid-scale topographic variations may play a predominant role. Research on the integration of subgrid-scale topography into macrosocopic shallow water models is thus needed. Upscaling of friction/turbulence-induced head loss terms is also a subject for research, with a number of competing approaches available from the literature [42], [43], [55], [58].
- Upscaling of transport processes. The upscaling of surface pollutant transport processes in the urban environment has not been addressed so far in the literature. Free surface flows in urban areas are characterized by strongly variable (in both time and space) flow fields. Dead/swirling zones have been shown to play a predominant role in the upscaling of the flow equations [43], [55]. Their role is expected to be even stronger in the upscaling of contaminant transport. While numerical experiments indicate that the microscopic hydrodynamic time scales are small compared to the macroscopic time scales, theoretical considerations indicate that this may not be the case with scalar transport. Trapping phenomena at the microscopic scale are well-known to be upscaled in the form of fractional dynamics models in the long time limit [46], [52]. The difficulty in the present research is that upscaling is not sought only for the long time limit but also for all time scales. Fractional dynamics will thus probably not suffice to a proper upscaling of the transport equations at all time scales.

3.2.3. Numerical platform

As a long term objective, the team shall create a common architecture for existing codes, and also the future codes developed by the project members, to offer a simplified management of various evolutions and a single and well documented tool for our partners. It will aim to be self-contained including pre and post-processing tools (efficient meshing approaches, GMT and VTK libraries), but must of course also be opened to user's suggestions, and account for existing tools inside and outside Inria. This numerical platform will be dedicated to the simulation of all the phenomena of interest, including flow propagation, sediment evolution, model

coupling on large scales, from deep water to the shoreline, including swell propagation, shoaling, breaking and run-up. This numerical platform clearly aims at becoming a reference software in the community. It should be used to **develop a specific test case** around Montpellier which embeds many processes and their mutual interactions: from the *Camargue* (where the Rhône river flows into the Mediterranean sea) to the *Étang de Thau* (a wide lagoon where shellfishes are plentiful), **all the processes studied in the project occur in a 100km wide region**, including of course the various hydrodynamics regimes (from the deep sea to the shoaling, surf and swash zones) and crucial morphodynamic issues (*e.g.* in the town of Sete).

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 elastoacoustic and electroseismic 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 Electroseismology.
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 [90], [68]. 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 [71]. 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 [80], 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 [65], 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 [95]. 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. Electroseismic 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 [88], 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 [70]. 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. Electroseismology 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 Electroseismology [97], [98], [85], [86]. 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 [72]. 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 [66], [67], 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 [69], [76], [75], [73], [77]. 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 [74] 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 [89], [94], [92]. 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.enseeiht.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 elastodynamic 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 [81], [78], [96], [91], [84] 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 [87].

Recently, Dolean *et al* [83] 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 [79]. 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 visco-elastic 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. The load-balancing leads then to the issue of mesh partitioning. Main mesh partitioners are very efficient for the coupling of different discretizations in space but to the best of our knowledge, the case of non-uniform time discretization has never been addressed. The study of meshes being out of the scopes of Magique-3D, we will collaborate with experts on mesh partitioning. We get already on to François Pellegrini who is the principal investigator of Scotch (http://www.labri.fr/perso/pelegrin/scotch) and permanent member of the team project Bacchus (Inria Bordeaux Sud Ouest Research Center).

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.

SAGE Project-Team

3. Research Program

3.1. Numerical algorithms and high performance computing

Linear algebra is at the kernel of most scientific applications, in particular in physical or chemical engineering. For example, steady-state flow simulations in porous media are discretized in space and lead to a large sparse linear system. The target size is 10^7 in 2D and 10^{10} in 3D. For transient models such as diffusion, the objective is to solve about 10^4 linear systems for each simulation. Memory requirements are of the order of Giga-bytes in 2D and Tera-bytes in 3D. CPU times are of the order of several hours to several days. Several methods and solvers exist for large sparse linear systems. They can be divided into three classes: direct, iterative or semi-iterative. Direct methods are highly efficient but require a large memory space and a rapidly increasing computational time. Iterative methods of Krylov type require less memory but need a scalable preconditioner to remain competitive. Iterative methods of multigrid type are efficient and scalable, used by themselves or as preconditioners, with a linear complexity for elliptic or parabolic problems but they are not so efficient for hyperbolic problems. Semi-iterative methods such as subdomain methods are hybrid direct/iterative methods which can be good tradeoffs. The convergence of iterative and semi-iterative methods and the accuracy of the results depend on the condition number which can blow up at large scale. 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.

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. Complex models are in general coupled multi-physics models. 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.

Approximation in Krylov subspace is in the core of the team activity since it provides efficient iterative solvers for linear systems and eigenvalue problems as well. The later are encountered in many fields and they include the singular value problem which is especially useful when solving ill posed inverse problems.

3.2. Numerical models applied to hydrogeology and physics

The team Sage is strongly involved in numerical models for hydrogeology and physics. There are many scientific challenges in the area of groundwater simulations. This interdisciplinary research is very fruitful with cross-fertilizing subjects. For example, high performance simulations were very helpful for finding out the asymptotic behaviour of the plume of solute transported by advection-dispersion. Numerical models are necessary to understand flow transfer in fractured media.

The team develops stochastic models for groundware simulations. Numerical models must then include Uncertainty Quantification methods, spatial and time discretization. Then, the discrete problems must be solved with efficient algorithms. The team develops parallel algorithms for complex numerical simulations and conducts performance analysis. Another challenge is to run multiparametric simulations. They can be multiple samples of a non intrusive Uncertainty Quantification method, or multiple samples of a stochastic method for inverse problems, or multiple samples for studying the sensitivity to a given model parameter. Thus these simulations are more or less independent and are well-suited to grid computing but each simulation requires powerful CPU and memory resources.

A strong commitment of the team is to develop the scientific software platform H2OLab for numerical simulations in heterogeneous hydrogeology.

SERENA Team

3. Research Program

3.1. Multiphysics coupling and domain decomposition

Within our project, we start from the conception and analysis of *models* based on *partial differential equations*. Already at the PDE level, we address the question of *coupling* of different models; examples may be 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 request to satisfy. At the modeling stage, we also plan to 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.

We also concentrate an important effort on the development and analysis of efficient solvers for the systems of nonlinear algebraic equations. We have already in the past developed *Newton–Krylov solvers*, with a particular impact on *parallelization* that we achieve via *domain decomposition*. Here we specialize on Robin boundary conditions, where an optimized choice of the parameter has already shown speed-ups in orders of magnitude in terms of the number of the iterations of the domain decomposition algorithm in question. A novel feature is the use of such algorithms in time-dependent problems in *space-time* domain decomposition which allows the use of different time steps in different parts of the computational domain. This is particularly useful in porous media applications, where the amount of diffusion (permeability) varies abruptly, so that the evolution speeds vary importantly and call for adapted localized time stepping. Our other novel theme are *Newton–multigrid solvers*, where the geometric multigrid solver ingredients are *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* on *each iteration step*, a very welcome feature in most of the target applications. The solver itself is then *steered adaptively* at each execution step by an a posteriori error estimate.

3.2. Structure-preserving discretizations and discrete element methods

Our second important activity is the design of *numerical methods* for the devised partial differential equation 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 *non-matching interfaces*, and we develop them in response to a high demand from our industrial partners. 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 developed schemes, we study the existence, uniqueness, and stability questions, and derive a priori convergence estimates. Our special interest is in higher-order methods, almost nonexisting these days even on the theoretical side. Albeit their use in practice may not be immediate, we believe that they represent the future generation of numerical methods for industrial simulations. These developments are and will be undertaken by Alexandre Ern and Laurent Monasse from ENPC who shall enlarge SERENA in the joint Inria–ENPC team shortly.

3.3. Reliability by a posteriori error control and safe and correct programming

The last part of our theoretical efforts goes towards the certification of the results obtained at the end of the numerical simulation. Here a key ingredient is the development of rigorous *a posteriori error estimates* that enable to estimate in a guaranteed 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 that coming from the modeling, from the discretization scheme, from the nonlinear (Newton) solver, and from the linear algebraic (Krylov, domain decomposition, multigrid) solver. A novel concept here is that of *local stopping criteria*, where all the error components are balanced locally within each computational mesh element. This naturally interconnects all parts of the numerical simulation process and gives rise to novel *fully adaptive algorithms*. We in particular aim to address the question of convergence of the new algorithms and justify their optimality, proving in particular guaranteed error reduction.

We also concentrate on the issue of computer implementation of scientific computing programs. The 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*. In our project, we explore the use of these tools to design generic solutions for the implementation of the considered class of scientific computing software. Our focus ranges from high-level programming via *functional programming* with OCaml, and safe and easy parallelism via *skeleton parallel programming* with Sklml, to proof of correctness of numerical algorithms and programs via *mechanical proofs* with Coq.

STEEP 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. URBANSIM), 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. A second LUTI model that will be considered in the near future, within the CITiES project, is UrbanSim [0]. Whereas TRANUS aggregates over e.g. entire population or housing categories, UrbanSim takes a micro-simulation approach, modeling and simulating choices made at the level of individual households, businesses, and jobs, for instance, and it operates on a finer geographic scale than TRANUS.

[0] http://www.urbansim.org

On the other hand, the scientific domains related to eco-system 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 and material flow analysis only proposes statistical models based on more or less simple data correlations. The eco-system service community has been using statical 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 will work in particular on a land use/ land cover change (LUCC) modelling environments (LCM from Clark labs [0], and 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

[0] http://www.clarklabs.org/products/Land-Change-Modeler-Overview.cfm
[0] http://www.csr.ufmg.br/dinamica/

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

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 divising 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. All these uses of SA will be considered in our research.

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 will have to be circumvented e.g. difficulties arising from dependencies of input variables, variables that obey a spatial organization, or switch inputs. We will 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 will 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 will work towards using sensitivity indices based on functional analysis of variance, which will allow us to compare the influence of various inputs on the indicators. For example it will 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 into account relativistic particles, which lead to consider the relativistic Vlasov equation, but generally, tokamak plasmas are supposed to be non relativistic. The particles distribution function depends on seven variables (three for space, three for 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 prohibitive amount of calculations. It is then necessary to develop models where the cyclotron frequency tends to infinity 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. Those models require averaging of the acting fields during a rotation period along the trajectories of the particles. This averaging is called the gyroaverage and requires specific discretizations.

The tokamak and its magnetics fields present a very particular geometry. Some authors have proposed to return to the intrinsic geometrical versions of the Vlasov-Maxwell system in order to build better gyrokinetic models and adapted numerical schemes. This implies the use of sophisticated tools of differential geometry: differential forms, symplectic manifolds, and hamiltonian geometry.

In addition to theoretical modeling tools, it is necessary to develop numerical schemes adapted to kinetic and gyrokinetic 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 where 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. Implemented Two-Scale Numerical Methods (for instance by Frénod et al. [30]) are 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 to 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) with the property of being able to 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. [28].

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 longstanding 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 introduce 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) [38] for the simulation of core turbulence that has been developed at CEA Cadarache in collaboration with our team [31]. Although GYSELA was carefully developed to be conservative at lowest order, it is not exactly conservative, which might be an issue when the simulation is under-resolved, which always happens in turbulence simulations due to the formation of vortices which roll up.

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 the method 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 longstanding experience in PIC methods and we started implement them in Selalib. An important aspect is to adapt the method to new multicore computers. See the work by Crestetto and Helluy [27].

3.2. 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 MagnetoHydroDynamic (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. Fine collision models are mainly investigated by other partners of the IPL (Inria Project Lab) FRATRES. In our approach we do not assume that the collisions are strong, but rather try to adapt the representation of the distribution function according to its shape, keeping the kinetic effects. 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). Recently, a team at the Plasma Physics Institut (IPP) in Garching has carried out a statistical analysis of the 5D distribution functions obtained from gyrokinetic tokamak simulations [32]. They discovered that the fluctuations are much higher in the space directions than in the velocity directions (see Figure 1).

This indicates that the approximation of the distribution function could require fewer data while still achieving a good representation, even in the collisionless regime.

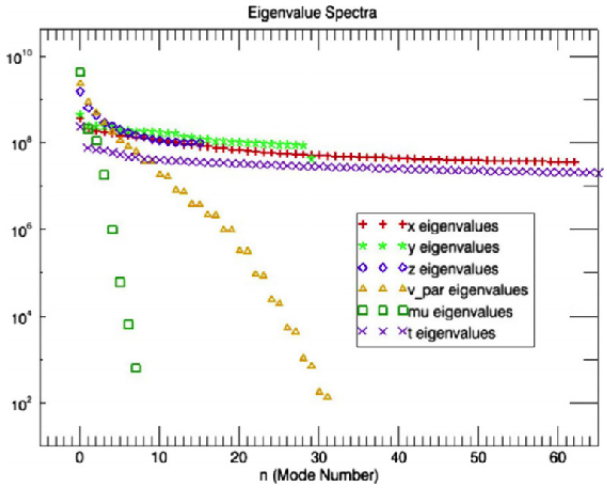


Figure 1. Space and velocity fluctuations spectra (from [32])

Our approach is different from the fluid approximation. In what follows we call this the "reduced model" approach. 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 solution of a first order hyperbolic system.

Several approaches are possible: waterbag approximations, velocity space transforms, *etc.*

3.2.1. Velocity space transformations

An experiment made in the 60's [35] exhibits in a spectacular way the reversible nature of the Vlasov equations. When two perturbations are applied to a plasma at different times, at first the plasma seems to damp and reach an equilibrium. But the information of the perturbations is still here and "hidden" in the high frequency microscopic oscillations of the distribution function. At a later time a resonance occurs and the plasma produces an echo. The time at which the echo occurs can be computed (see Villani [0], page 74). The fine mathematical study of this phenomenon allowed C. Villani and C. Mouhot to prove their famous result on the rigorous nonlinear Landau damping [37].

More practically, this experiment and its theoretical framework show that it is interesting to represent the distribution function by an expansion on an orthonormal basis of oscillating functions in the velocity variables. This representation allows a better control of the energy transfer between the low frequencies and the high frequencies in the velocity direction, and thus provides more relevant numerical methods. This kind of approach is studied for instance by Eliasson in [29] with the Fourier expansion.

[0] Landau damping. CEMRACS 2010 lectures. http://smai.emath.fr/cemracs/cemracs10/PROJ/Villani-lectures.pdf

In long time scales, filamentation phenomena result in high frequency oscillations in velocity space that numerical schemes cannot resolve. For stability purposes, most numerical schemes contain dissipation mechanisms that may affect the precision of the finest oscillations that could be resolved.

3.2.2. Adaptive modeling

Another trend in scientific computing is to optimize the computation time through adaptive modeling. This approach consists in applying the more efficient model locally, in the computational domain, according to an error indicator. In tokamak simulations, this kind of approach could be very efficient, if we are able to choose locally the best intermediate kinetic-fluid model as the computation runs. This field of research is very promising. It requires developing a clever hierarchy of models, rigorous error indicators, versatile software architecture, and algorithms adapted to new multicore computers.

3.2.3. 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 [39], [25], [34], [33].

3.3. Electromagnetic solvers

A 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 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 [22] (and included references).

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 [86]. 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 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 an activation wave that rapidly propagates through the tissue. The spatio-temporal pattern of this propagation is related both to the function of the cellular membrane and to the structural organisation of the cells into tissues. Cardiac arrythmias originate from malfunctions in this process. The field of cardiac electrophysiology studies the multiscale organisation 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.

Several improvements of current models of the propagation of the action potential will be developped, based on previous work and on the data available at the LIRYC:

- Enrichment of the current monodomain and bidomain models 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 want to develop model that couples 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 and check that they are reliable experimental tools.

3.2. Simplified models and inverse problems

The medical and clinical exploration of the electrical signals is based on accurate reconstruction of the typical patterns of propagation of the action potential. The correct detection of these complex patterns by non-invasive electrical imaging techniques has to be developed. Both problems involve solving inverse problems that cannot be addressed with the more compex 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 in the cardiac diagnosis. Although widely studied during the last years, it remains a challenge for the scientific community. In many cases the quality of reconstructed electrical potential is not sufficiently accurate. The methods used consist in solving the Laplace equation on the volume delimited by the body surface and the epicardial surface. We want to

- Study in depth the dependance of this inverse problem inhomogeneities in the torso, conductivity values, the geometry, electrode placements...
- Improve the solution to the inverse problem be using new regularization strategies 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 conside the follwong strategies:

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

Additionaly, 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] and [6] and [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 preconditionning 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 parallellization, 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 cardiac physiology are based on the approximation of a perfect muscle using homogenisation method. However, due to the age and due to some cardiomyopathies, the cellular structure of the tissue changes. These modifications 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 numerically explore the mechanisms of these diseases.

The problem is that literature on this type of model is still very poor and existing models are bidimensionels or limited to idealised geometries. We propose an approach in opposition with the usual homogenisation way. We want to describe the muscle as a system of three-dimensional cells, whose dynamics is given by the modeling of ion fluxes across cell membranes in equilibrium with the electrostatic potentials in the intracellular and extracellular environments.

The goals are to design, analyse, and explore numerically a model of cardiac electrophysiology at a level of discretisation of about $1\mu m$ (that means 10 to 100 times smaller than the size of cardiomyocytes), develop model and its numerical discretisation, define realistic geometries or actual cells.

Issues are scale simulations for thousands of cores to take into account thousands or tens of thousands cells. For this, a hybrid parallelism approach OpenMP and MPI will be considered.

DRACULA Project-Team

3. Research Program

3.1. Cell dynamics

We model dynamics of cell populations with two approaches, dissipative particle dynamics (DPD) and partial differential equations (PDE) of continuum mechanics. DPD is a relatively new method developed from molecular dynamics approach largely used in statistical physics. Particles in DPD do not necessarily correspond to atoms or molecules as in molecular dynamics. These can be mesoscopic particles. Thus, we describe in this approach a system of particles. In the simplest case where each particle is a sphere, they are characterized by their positions and velocities. The motion of particles is determined by Newton's second law (see Figure 1).

In our case, particles correspond to biological cells. The specific feature of this case in comparison with the conventional DPD is that cells can divide (proliferation), change their type (differentiation) and die by apoptosis or necrosis. Moreover, they interact with each other and with the extra-cellular matrix not only mechanically but also chemically. They can exchange signals, they can be influenced by various substances (growth factors, hormones, nutrients) coming from the extra-cellular matrix and, eventually, from other organs.

Distribution of the concentrations of bio-chemical substances in the extra-cellular matrix will be described by the diffusion equation with or without convective terms and with source and/or sink terms describing their production or consumption by cells. Thus we arrive to a coupled DPD-PDE model.

Cell behaviour (proliferation, differentiation, apoptosis) is determined by intra-cellular regulatory networks, which can be influenced by external signals. Intra-cellular regulatory networks (proteins controlling the cell cycle) can be described by systems of ordinary differential equations (ODE). Hence we obtain DPD-PDE-ODE models describing different levels of cell dynamics (see Figure 1). It is important to emphasize that the ODE systems are associated to each cell and they can depend on the cell environment (extra-cellular matrix and surrounding cells).

3.2. From particle dynamics to continuum mechanics

DPD is well adapted to describe biological cells. However, it is a very time consuming method which becomes difficult to use if the number of particles exceeds the order of 10^5-10^6 (unless distributed computing is used). On the other hand, PDEs of continuum mechanics are essentially more efficient for numerical simulations. Moreover, they can be studied by analytical methods which have a crucial importance for the understanding of relatively simple test cases. Thus we need to address the question about the relation between DPD and PDE. The difficulty follows already from the fact that molecular dynamics with the Lennard-Jones potential can describe very different media, including fluids (compressible, incompressible, non-Newtonian, and so on) and solids (elastic, elasto-plastic, and so on). Introduction of dissipative terms in the DPD models can help to justify the transition to a continuous medium because each medium has a specific to it law of dissipation. Our first results [40] show the correspondence between a DPD model and Darcy's law describing fluid motion in a porous medium. However, we cannot expect a rigorous justification in the general case and we will have to carry out numerical comparison of the two approaches.

An interesting approach is related to hybrid models where PDEs of continuum mechanics are considered in the most part of the domain, where we do not need a microscopical description, while DPD in some particular regions are required to consider individual cells.

3.3. PDE 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. These are more traditional models [41] with properties that depend on the particular problem under consideration and with many open questions, both from the point of view of their mathematical properties and for applications. In particular we are interested in the spreading of cell populations which describes the development of leukemia in the bone marrow and many other biological phenomena (solid tumors, morphogenesis, atherosclerosis, and so on). From the mathematical point of view, these are reaction-diffusion waves, intensively studied in relation with various biological problems. We will continue our studies of wave speed, stability, nonlinear dynamics and pattern formation. From the mathematical point of view, these are elliptic and parabolic problems in bounded or unbounded domains, and integro-differential equations. We will investigate the properties of the corresponding linear and nonlinear operators (Fredholm property, solvability conditions, spectrum, and so on). Theoretical investigations of reaction-diffusion-convection models will be accompanied by numerical simulations and will be applied to study hematopoiesis.

Hyperbolic problems are also of importance when describing cell population dynamics ([46], [48]), and they proved effective in hematopoiesis modelling ([35], [36], [38]). They are structured transport partial differential equations, in which the structure is a characteristic of the considered population, for instance age, size, maturity, protein concentration, etc. The transport, or movement in the structure space, simulates the progression of the structure variable, growth, maturation, protein synthesis, etc. Several questions are still open in the study of transport PDE, yet we will continue our analysis of these equations by focusing in particular on the asymptotic behaviour of the system (stability, bifurcation, oscillations) and numerical simulations of nonlocal transport PDE.

The use of age structure often leads to a reduction (by integration over the age variable) to nonlocal problems [48]. The nonlocality can be either in the structure variable or in the time variable [35]. In particular, when coefficients of an age-structured PDE are not supposed to depend on the age variable, this reduction leads to delay differential equations.

3.4. Delay differential Equations

Delay differential equations (DDEs) 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. Because these processes can take a certain time, the system depends on an essential way of its past state, and can be modelled by DDEs.

We explain hereafter how delays can appear in hematopoietic models. Based on biological aspects, we can divide hematopoietic cell populations into many compartments. We basically consider two different cell populations, one composed with immature cells, and the other one made of mature cells. Immature cells are separated in many stages (primitive stem cells, progenitors and precursors, for example) and each stage is composed with two sub-populations, resting (G0) and proliferating cells. On the opposite, mature cells are known to proliferate without going into the resting compartment. Usually, to describe the dynamic of these multi-compartment cell populations, transport equations (hyperbolic PDEs) are used. Structure variables are age and discrete maturity. In each proliferating compartment, cell count is controlled by apoptosis (programmed cell death), and in the other compartments, cells can be eliminated only by necrosis (accidental cell death). Transitions between the compartments are modelled through boundary conditions. In order to reduce the complexity of the system and due to some lack of information, no dependence of the coefficients on cell age is assumed. Hence, the system can be integrated over the age variable and thus, by using the method of characteristics and the boundary conditions, the model reduces to a system of DDEs, with several delays.

Leaving all continuous structures, DDEs appear well adapted to us to describe the dynamics of cell populations. They offer good tools to study the behaviour of the systems. The main investigation of DDEs are the effect of perturbations of the parameters, as cell cycle duration, apoptosis, differentiation, self-renewal, and re-introduction from quiescent to proliferating phase, on the behaviour of the system, in relation for instance with some hematological disorders [42].

M3DISIM 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 [5]. 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 [6].

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

At small spatial scales, or at spatial scales of individual matter components, where heterogeneities in the medium occur, agent-based models are developed ([0], [75], Dirk Drasdo's former associate team QUANTISS). Another approach, that is considered in the project-team MAMBA consists in considering gene expression at the individual level by stochastic processes [0] or by ordinary differential equations [0], or by a mixed representation of Markov processes and ordinary differential equations [0], the outputs of which quantify focused aspects of biological variability in a population of individuals (cells) under study.

Both these approaches complement the partial differential equation models considered on scales at which averages over the individual components behave sufficiently smoothly. Investigating the links between these models through scales is also part of our research [0]. Moreover, in order to quantitatively assess the adequacy between the biological phenomena we study and the mathematical models we use, we also develop inverse problem methods.

3.2. PDE analysis and simulation

PDEs arise at several levels of our models. Parabolic equations can be used for large cell populations and also for intracellular spatio-temporal dynamics of proteins and their messenger RNAs in gene regulatory networks, transport equations are used for protein aggregation / fragmentation models and for the cell division cycle in age-structured models of proliferating cell populations. Existence, uniqueness and asymptotic behaviour of solutions have been studied [0] [67] [65]. Other equations, of the integro-differential type, dedicated to describing the Darwinian evolution of a cell population according to a phenotypic trait, allowing exchanges with the environment, genetic mutations and reversible epigenetic modifications, are also used [82], [83], [81] [23]. Through multiscale analysis, they can be related to stochastic and free boundary models used in cancer modelling.

Inverse problems

When studying biological populations (usually cells or big molecules) using PDE models, identification of the functions and parameters that govern the dynamics of a model may be achieved to a certain extent by statistics performed on individuals to reconstruct the probability distribution of their relevant characteristics in the population they constitute, but quantitative observations at the individual level (e.g., fluorescence in single cells [63] or size/age tracking [89]) require sophisticated techniques and are most often difficult to obtain. Relying on the accuracy of a PDE model to describe the population dynamics, inverse problem methods offer a tractable alternative in model identification, and they are presently an active theme of research in MAMBA. Following previous studies [70], [71], some combining statistical and deterministic approaches [69] with application to raw experimental data [13], we plan to develop our methods to new structured-population models (or stochastic fragmentation processes as in [68]), useful for other types of data or populations (e.g. size/age tracking, polymer length distribution, fluorescence in single cells).

[0] Drasdo, Hoehme, Block, *J. Stat. Phys.*, 2007

[0] as in M. Sturrock et al., Spatial stochastic modelling of the Hes1 gene regulatory network: intrinsic noise can explain heterogeneity in embryonic stem cell differentiation, *Journal of The Royal Society Interface*, 2013

[0] as in A. Friedman et al, Asymptotic limit in a cell differentiation model with consideration of transcription, *J. Diff. Eq.*, 2012

[0] as in R. Yvinec et al., Adiabatic reduction of stochastic gene expression with jump Markov processes, *J. Math. Biol.*, 2013.

[0] H. Byrne and D. Drasdo, Individual-based and continuum models of growing cell populations: a comparison, *J. Math. Biol*, 2009

[0] B. Perthame, Transport equations in biology, Springer, 2007

3.3. Stochastic and agent-based models

The link between stochastic processes and kinetic equations is a domain already present in our research [0] [69] and that we plan to develop further. They can be viewed either as complementary approaches, useful to take into account different scales (smaller scales for stochastic models, larger scales for mean-field limits), or even as two different viewpoints on the same problem [68], enriching each other. Neuroscience is a domain where this is particularly true because noise contributes significantly to the activity of neurons; this is the case of networks where mean field limits are derived from stochastic individual-based models and lead to fundamental questions on the well-posedness and behaviours of the system [0]. One strength and originality of our project is our close connection and collaboration not only with probability theorists but also with statisticians, who provide us with efficient help in the identification of our model parameters.

Agent-based systems consider each component individually. For example, in multi-cellular system modelling, the basic unit is the cell, and each cell is considered [72], [29]. This approach has advantages if the population of cells reveals inhomogeneities on small spatial scales as it occurs if organ architecture is represented [75], or if the number of cells in a particular state is small. Different approaches have been used to model cellular agents in multi-cellular systems in space, roughly divided in lattice models (e.g. [88]) and in lattice-free (or off-lattice) models, in which the position [72], [74] or even the shape (e.g. [29]) of the cell can change gradually.

The dynamics of cells in lattice-based models is usually described by rules chosen to mimic the behaviour of a cell including its physical behavior. The advantage of this approach is that it is simpler and that simulation times for a given number of cells are shorter than in lattice-free models. In contrast, most lattice-free models attempt to parameterise cells by measurable values with a direct physical or biological meaning, hence allowing identification of physiologically meaningful parameter ranges. This improves model simulation feasibility, since parameter sensitivity analyses in simulations shows significant improvements when a high dimensional parameter space can be reduced. It also facilitates the development of systematic systems biology and systems medicine strategies to identify mechanisms underlying complex tissue organisation processes ([29], [73]).

Moreover, it is straightforward to include relevant signal transduction and metabolic pathways in each cell within the framework of agent-based models, which is a key advantage in the present times, as the interplay of components at many levels is more and more precisely studied [93].

[0] H. Byrne and D. Drasdo, Individual-based and continuum models of growing cell populations: a comparison, *J. Math. Biol*, 2009

[0] Cáceres, Carrillo, Perthame *J. Math. Neurosci.* 2011; Pakdaman, Perthame, Salort *Nonlinearity* 2010

MODEMIC Project-Team

3. Research Program

3.1. Modeling and simulating microbial ecosystems

The chemostat model is quite popular in microbiology and bioprocess engineering [58], [60]. Although the wording "chemostat" refers to the experimental apparatus dedicated to continuous culture, invented in the fifties by Monod and Novick & Szilard, the chemostat model often serves as a mathematical representation of biotic/abiotic interactions in more general (industrial or natural) frameworks of microbial ecology. The team carries a significant activity about generalizations and extensions of the classical model (see Equation (1) and Section 3.1.1) which assumes that the sizes of the populations are large and that the biomass can be faithfully represented as a set of deterministic continuous variables.

However recent observations tools based notably on molecular biology (e.g. molecular fingerprints) allow to distinguish much more precisely than in the past the internal composition of biomass. In particular, it has been reported by biologists that minority species could play an important role during transients (in the initialization phase of bio-processes or when the ecosystem is recovering from disturbances), that cannot be satisfactorily explained by the above deterministic models because the size of those populations could be too small for these models to be valid.

Therefore, we are studying extension of the classical model that could integrate stochastic/continuous macroscopic aspects, or microscopic/discrete aspects (in terms of population size or even with explicit individually based representation of the bacteria), as well as hybrid representations. One important question is the links between these chemostat models (see Section 3.1.2).

3.1.1. About the chemostat model

The classical mathematical chemostat model:

$$\begin{aligned} \dot{s} &= -\sum_{j=1}^{n} \frac{1}{y_j} \mu_j(s)\, x_j + D\,(s_{in} - s) \\ \dot{x}_i &= \mu_i(s)\, x_i - D\, x_i \qquad (i = 1 \cdots n) \end{aligned} \tag{19}$$

for n species in concentrations x_i competing for a substrate in concentration s, leads to the so-called "Competitive Exclusion Principle", that states that generically no more species than limiting resources can survive on a long term [59]. Apart some very precise laboratory experiments that have validated this principle, such an exclusion is rarely observed in practice.

Several possible improvements of the model (1) need to be investigated, related to biologists' knowledge and observations, in order to provide better interpretations and predictive tools. Various extensions have already been studied in the literature (e.g. crowding effect, inter-specific interactions, predating, spatialization, time-varying inputs...) to which the team has also contributed. This is always an active research topic in bio-mathematics and theoretical ecology, and several questions remains open or unclear, although numerical simulations guide the results to be proven.

Thanks to the proximity with biologists, the team is in position to propose new extensions relevant for experiments or processes conducted among the application partners. Among them, we can mention: intra and inter-specific interactions terms between microbial species; distinction between planktonic and attached biomass; effects of interconnected vessels; consideration of maintenance or variable yield in the growth reactions; coupling with membrane fouling mechanisms.

Our philosophy is to study how complex or not very well known mechanisms could be represented satisfactorily by simple models. It often happens that these mechanisms have different time scales (for instance the flocculation of bacteria is expected to be much faster than the biomass growth), and we typically use singular perturbations

techniques to produce reduced models.

3.1.2. Stochastic and multi-scale models

Comparatively to deterministic differential equations models, quite few stochastic models of microbial growth have been worked out in the literature. Nonetheless, numerous problems could benefit from such an approach (dynamics with small population sizes, persistence and extinction, random environments...).

For example, the need to clarify the role of minority species conducts to revisit thoroughly the chemostat model at a microscopic level, with birth and death or pure jump processes, and to investigate which kind of continuous models it raises at a macroscopic scale. For this purpose, we consider the general framework of Markov processes [57].

It also happens that minority species cohabit with other populations of much larger size, or fluctuate with time between small and large sizes. There is consequently a need to build new "hybrid" models, that have individual-based and deterministic continuous parts at the same time. The persistence (temporarily or not) of minority species on the long term is quite a new questioning spread in several applications domains at the Inra Institute.

Continuous cultures of micro-organisms often face random abiotic environments, that could be considered as random switching between favorable or unfavorable environments. This feature could lead to non-intuitive behaviors in long run, concerning persistence or extinction of populations. We consider here the framework of piece-wise deterministic Markov processes [55].

3.1.3. Computer simulation

The simulation of dynamical models of microbial ecosystems with the features described in Section 3.1.2 raises specific and original algorithmic problems:

- simultaneous presence in the same algorithms of both continuous variables (concentration of chemicals or very large populations) and discrete (when the population has a very small number of individuals),
- simultaneous presence in the same algorithms of stochastic aspects (for demographic and environmental noises) and deterministic ones (when the previous noises are negligible at macroscopic scales)
- use of individual-based models (IBM) (usually for small population sizes).

We believe that these questions must be addressed in a rigorous mathematical framework and that their solutions as efficient algorithms are a formidable scientific challenge.

3.2. Identification and control

3.2.1. Models identification and state estimation

Growth kinetics is usually one of the crucial ingredients in the modeling of microbial growth. Although the specific growth rate functions and their parameters can be identified in pure cultures (and can be estimated with accuracy in laboratory experiments), it is often an issue to extrapolate this knowledge in industrial setup or in mixed cultures. The parameters of these functions could change with their chemical and physical environment, and species interactions could inhibit or promote a strain that is expected to dominate or to be dominated in an multi-species ecosystem. Moreover, we need to estimate the state variables of the models.

We aim at developing effective tools for the on-line reconstruction of growth curves (and of their parameters) and/or state variables, along with the characteristics of microbial ecosystems:

- It is not always possible to drive a biological system for exploring a large subset of the state space, and open-loop dynamics could be unstable when far from locally stable equilibria (for instance under inhibition growth).
- The number of functional groups of species and the nature of their interactions (competition, mutualism, neutral) are not always known a priori and need to be estimated.

We look for observers or filters based methods (or alternatives), as well as estimation procedures, with the typical difficulty that for biological systems and their outputs it is rarely straightforward to write the models into a canonical observation form. However, our objective is to obtain an adjustable or guaranteed speed of convergence of the estimators.

3.2.2. Optimal design and control

For practitioners, an expected outcome of the models is to bring improvements in the design and real-time operation of the processes. This naturally leads to mathematical formulations of optimization, stabilizing control or optimal control problems. We distinguish two families of problems:

- *Process design and control within an industrial setup.* Typically one aims at obtaining small residence times for given input-output performances and (globally) stable processes. The design questions consist in studying on the models if particular interconnections and fill strategies allow to obtain significant gains. The specificity of the models and the inputs constraints can lead to systems that are not locally controllable, and thus the classical linearizing techniques do not work. This leaves open some problems for the determination of globally stabilizing feedback or optimal syntheses.
- *Design and control for resource preservation in natural environments (such as lakes, soil bio-remediation...).* Here, the spatial heterogeneity of the resource might be complex and/or not well known. We look for sparse spatial representations in order to apply finite dimensional tools of state-space systems.

In both cases, one faces model uncertainty and partial measurements that often require to couple the techniques developed in Section 3.2.1 .

Monc Team

3. Research Program

3.1. Introduction

We address the problem of cancer modeling through 3 axis.

- Axis 1: Tumor modeling for patient-specific simulations.
- Axis 2: Bio-physical modeling for personalized therapies.
- Axis 3: Quantitative cancer modeling for biological and preclinical studies.

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 give information to the clinicians in order to improve the decision process. It is mainly useful in the case of a relapse or for metastatic diseases.

The second axis aims at modeling the 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 are not 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 adopt a strategy in order 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 follow 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 gliomas). If we had a reliable patient specific model of tumor growth with or without treatment, it could have different uses for the patient follow-up.

- Case without treatment: the evaluation of the growth of the tumor would provide 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 for the planification of the intervention. For tumor with very slow growth, 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 forecast, this can mean that some events have occurred at the biological level. It can be the apparition of an agressive phenotype, 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 effect 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 a coupling of gradient methods and Monte-Carlo type 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 O.D.E. or a P.D.E that describes the evolution of the growth rate as a combination of sigmoidal 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 gliomas 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 a 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 done 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.

3.3. Axis 2: Bio-physical modeling for personalized therapies

In this axis, we investigate locoregional therapies such as radiotherapy, irreversible electroporation. Electroporation consists of an increase of the membrane permeability of cells due to the delivery of high voltage pulses. This phenomenon can be transient (reversible) or irreversible. This is a non-thermal phenomenon. (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 will focus on two particular approaches: some aspects of radiotherapies and electro-chemotherapy. This choice is motivated by some pragmatic reasons: we already have collaborations with physicians on these therapies. Other treatments could be probably tackled in the same spirit, 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 the healthy tissue. 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 is being further 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 a starting collaboration with Institut Bergonié. The response is evaluated from image markers (*e.g.* using texture information) or with a mathematical model (another collaboration is also ongoing with LATIM team in Brest on response of colorectal tumors to RT using PET scans). 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 since a few decades but it is still not well understood, therefore we address the modeling two different purposes:
  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 bioeletromagneticians 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 (LMPEO) 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...), on 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 theses 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.

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 potentiate 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 recurrent technical challenge that we need to address 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 iterative 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 [38], [37], and a nD, multiscale in space, system of weakly-coupled non conservative transport equations that underlies our approach of gonadal cell dynamics [39],[7]. Most 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 [8]. 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 [4]. 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 [41] 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.

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 $2n$D 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 [11], while associated control problems have been studied in the framework of hybrid optimal control [5]. 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 [3] and implemented in a problem-specific way on parallel architecture [1].

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 [42], [36]. 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 [14].

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 [13], [12] 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. Multiscale propagation phenomena in biology

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 bacteira, evoluation of age structures 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.

Multiscale modeling of propagation phenomena raises a lot of interest in several fields of application. This ranges from shock waves in kinetic equations (Boltzmann, BGK, etc...), bacterial chemotactic waves, selection-mutation models with spatial heterogeneities, evolution in age-structured population or subdiffusive processes.

Earlier works generally focused on numerical simulations, hydrodynamic limits to average over the microscopic variable, or specific models with only local features, not suitable for most of the relevant biological situations. Our contribution enables to derive the relevant features of propagation analytically, and far from the hydrodynamic regime for a wide range of models including nonlocal interaction terms.

Our recent understanding is closely related to the analysis of large deviations in multiscale dispersion equations (e.g. PDMP), for which we gave important contributions too in collaboration with E. Bouin (CEREMADE Dauphine), E. Grenier (Inria NUMED) and G. Nadin (Univ. Paris 6).

These advances are linked to the work of other Inria teams (MAMBA, DRACULA, BEAGLE), and collaborators in mathematics, physics and theoretical biology in France, Austria and UK.

3.1.2. Recent results

Vincent Calvez has focused on the modelling and analysis of propagation phenomena in structured populations. This includes chemotactic concentration waves, transport-reaction equations, coupling between ecological processes (reaction-diffusion) and evolutionary processes (selection of the fittest trait, adaptation), evolution of age structured poulations, and anomalous diffusion. As a main result, he could establish the existence of concentration waves of chemotactic bacteria E. coli in a fully coupled kinetic/reaction-diffusion system previously validated on experimental data.

In collaboration with a group of theoretical biologists at ISEM Montpellier (O. Ronce and O. Cotto), and J. Garnier (Univ. Savoie), Th. Lepoutre (Inria DRACULA), Th. Bourgeron (Inria NUMED) he has investigated quantitatively the maladaptation of an age-structured population in a changing environment. He has unravelled a striking phenomenon of severe maladaptation specific to age structure. This was observed on numerical simulations by biologists, but it has now a systematic mathematical comprehension.

He has also continued his work on the optimal control of monotone linear dynamical systems, using the Hamilton-Jacobi framework, and the weak KAM theory, in collaboration with P. Gabriel (UVSQ) and S. Gaubert (Inria MAXPLUS).

Alvaro Mateos Gonzalez has started his PhD on September 2014 under the supervision of Vincent Calvez, and Hugues Berry (BEAGLE), . He has already collaborated fruitfully with Thomas Lepoutre (DRACULA) and Hugues Berry to investigate the long-time asymptotics of a degenerate renewal equation. This is a first step towards the mathematical analysis of anomalous diffusion processes. In collaboration with P. Gabriel (UVSQ) and V. Calvez (Inria NUMED) he has investigated large deviations of heterogenous continuous time random walks.

3.1.3. Collaborations

- Mathematical description of bacterial chemotactic waves:
  - **N. Bournaveas** (Univ. Edinburgh), **V. Calvez** (ENS de Lyon, Inria NUMED) **B. Perthame** (Univ. Paris 6, Inria BANG), **Ch. Schmeiser** (Univ. Vienna), **N. Vauchelet**: design of the model, analysis of traveling waves, analysis of optimal strategies for bacterial foraging.
  - **J. Saragosti**, **V. Calvez** (ENS de Lyon, Inria NUMED), **A. Buguin**, **P. Silberzan** (Institut Curie, Paris): experiments, design of the model, identification of parameters.
- Transport-reaction waves and large deviations:
  - **E. Bouin**, **V. Calvez** (ENS de Lyon, Inria NUMED), **E. Grenier** (ENS de Lyon, Inria NUMED), **G. Nadin** (Univ. Paris 6)
- Selection-mutation models of invasive species:
  - **E. Bouin** (ENS de Lyon, Inria NUMED), **V. Calvez** (ENS de Lyon, Inria NUMED), **S. Mirrahimi** (Inst. Math. Toulouse): construction of traveling waves, asymptotic propagation of fronts,
  - **E. Bouin** (ENS de Lyon, Inria NUMED), **V. Calvez** (ENS de Lyon, Inria NUMED), **N. Meunier**, (Univ. Paris 5), **B. Perthame** (Univ. Paris 6, Inria Bang), **G. Raoul** (CEFE, Montpellier), **R. Voituriez** (Univ. Paris 6): formal analysis, derivation of various asymptotic regimes.
- Age-structured equations for anomalous diffusion processes, and evolution
  - **H. Berry** (Inria BEAGLE), **V. Calvez** (ENS de Lyon, Inria NUMED), **Th. Lepoutre** (Inria DRACULA), **P. Gabriel** (Univ. UVSQ), O. Ronce (ISEM Montpellier), O. Cotto (ISEM Montpellier), J. Garnier (Univ. Savoie).

3.2. Growth of biological tissues

3.2.1. Project-team positioning

The originality of our work is the derivation, analysis and numerical simulations of mathematical model for growing cells and tissues. This includes mechanical effects (growth induces a modification of the mechanical stresses) and biological effects (growth is potentially influenced by the mechanical forces).

This leads to innovative models, adapted to specific biological problems (*e.g.* suture formation, cell polarisation), but which share similar features. We perform linear stability analysis, and look for pattern formation issues (at least instability of the homogeneous state).

The biophysical literature of such models is large. We refer to the groups of Ben Amar (ENS Paris), Boudaoud (ENS de Lyon), Mahadevan (Harvard), etc.

Our team combines strong expertise in reaction-diffusion equations (V. Calvez) and mechanical models (P. Vigneaux). We develop linear stability analysis on evolving domains (due to growth) for coupled biomechanical systems.

Another direction of work is the mathematical analysis of classical tumor growth models. These continuous mechanics models are very close to classical equations like Euler or Navier Stokes equations in fluid mechanics. However they bring there own difficulties: Darcy law, multispecies equations, non newtonian dynamics (Bingham flows). Part of our work consist in deriving existence results and designing acute numerical schemes for these equations.

3.2.2. Recent results

We have worked on several biological issues. Cell polarisation is the main one. We first analyzed a nonlinear model proposed by theoretical physicists and biologists to describe spontaneous polarisation of the budding yeast *S. cerevisae*. The model assumes a dynamical transport of molecules in the cytoplasm. It is analogous to the Keller-Segel model for cell chemotaxis, except for the source of the transport flux. We developped nonlinear analysis and entropy methods to investigate pattern formation (Calvez et al 2012). We are currently validating the model on experimental data. The analysis of polarization of a single cell is a preliminary step before the study of mating in a population of yeast cells. In the mating phase, secretion of pheromones induces a dialogue between cells of opposite types.

We also derive realistic models for the growth of the fission yeast *S. pombe*. We proposed two models which couple growth and geometry of the cell. We aim to tackle the issue of pattern formation, and more specifically the instability of the spherical shape, leading to a rod shape. The mechanical coupling involves the distribution of microtubules in the cytoplasm, which bring material to the cell wall.

Over the evaluation period, Paul Vigneaux developped expertise in modelling and design of new numerical schemes for complex fluid models of the viscoplastic type. Associated materials are involved in a broad range of applications ranging from chemical industry to geophysical and biological materials. In the context of NUMED, this expertise is linked to the development of complex constitutive laws for cancer cell tissue. During the period, NUMED used mixed compressible/incompressible fluid model for tumor growth and viscoelastic fluid model. Viscoplastic is one of the other types of complex fluid model which is usable in the field. Mathematically, it involves variational inequalities and the need for specific numerical methods.

3.2.3. Collaborations

- **V. Calvez** (ENS de Lyon, Inria NUMED), **Th. Lepoutre** (Inria DRACULA), **N. Meunier**, (Univ. Paris 5), **N. Muller** (Univ. Paris 5), **P. Vigneaux** (ENS de Lyon, Inria NUMED): mathematical analysis of cell polarisation, numerical simulations
- **V. Calvez** (ENS de Lyon, Inria NUMED), **N. Meunier**, (Univ. Paris 5), **M. Piel**, (Institut Curie, Paris), **R. Voituriez** (Univ. Paris 6): biomechanical modeling of the growth of *S. pombe*
- **D. Bresch** (Univ. Chambéry), **V. Calvez** (ENS de Lyon, Inria NUMED), **R.H. Khonsari** (King's College London, CHU Nantes), **J. Olivier** (Univ. Aix-Marseille), **P. Vigneaux** (ENS de Lyon, Inria NUMED): modeling, analysis and simulations of suture formation.
- **Didier Bresch** (Univ Chambéry), **Benoit Desjardins**(Moma group): petrology.

ANR JCJC project "MODPOL", *Mathematical models for cell polarization*, led by Vincent Calvez (ENS de Lyon, CNRS, Inria NUMED).

3.3. Multiscale models in oncology

3.3.1. Project-team positioning

Since 15 years, the development of mathematical models in oncology has become a significant field of research throughout the world. Several groups of researchers in biomathematics have developed complex and multiscale continuous and discrete models to describe the pathological processes as well as the action of anticancer anti-cancer drugs. Many groups in US (e.g. Alexander Anderson's lab, Kristin Swansson's lab) and in Canada (e.g. Thomas Hillen, Gerda de Vries), quickly developed and published interesting modeling frameworks. The setup of European networks such as the Marie Curie research and training networks managed by Nicolas Bellomo and Luigi Preziosi constituted a solid and fertile ground for the development of new oncology models by teams of biomathematicians and in particular Zvia Agur (Israel), Philip Maini (UK), Helen Byrne (UK), Andreas Deutsch (Germany), or Miguel Herrero (Spain).

3.3.2. Results

We have worked on the development of a multiscale system for modeling the complexity of the cancer disease and generate new hypothesis on the use of anti-cancer drugs. This model relies on a multiscale formalism integrating a subcellular level integrating molecular interactions, a cell level (integrating the regulation of the cell cycle at the levels of individual cells) and a macroscopic level for describing the spatio-temporal dynamics of different types of tumor tissues (proliferating, hypoxic and necrotic). The model is thus composed by a set of partial differential equations (PDEs) integrating molecular network up to tissue dynamics using lax from fluid dynamic. This formalism is useful to investigate theoretically different cancer processes such as the angiogenesis and invasion. We have published several examples and case studies of the use of this model in particular, the action of phase-specific chemotherapies (Ribba, You et al. 2009), the use of anti-angiogenic drugs (Billy, Ribba et al. 2009) and their use in combination with chemotherapies (Lignet, Benzekry et al. 2013). This last work also integrates a model of the VEGF molecular pathway for proliferation and migration of endothelial cells in the context of cancer angiogenesis (Lignet, Calvez et al. 2013).

If these types of models present interesting framework to theoretically investigate biological hypothesis, they however present limitation due to their large number of parameters. In consequence, we decided to stop the development of the multiscale platform until exploration of alternative modeling strategies to deal with real data. We focus our interest on the use of mixed-effect modeling techniques as classically used in the field of pharmacokinetic and pharmacodynamics modeling. The general principal of this approach lies in the integration of several levels of variability in the model thus allowing for the simultaneous analysis of data in several individuals. Nowadays, complex algorithms allow for dealing with this problem when the model is composed by few ordinary differential equations (ODEs). However, no similar parameter estimation method is available for models defined as PDEs. In consequence, we decided: 1. To develop more simple models, based on systems of ODEs, assuming simplistic hypothesis of tumor growth and response to treatment but with a real focus on model ability to predict real data. 2. To work alone the development of parameter estimation methods for PDE models in oncology.

3.4. Parametrization of complex systems

3.4.1. Project-team positioning

We focus on a specific problem: the "population" parametrization of a complex system. More precisely, instead of trying to look for parameters in order to fit the available data for one patient, in many cases it is more pertinent to look for the distribution of the parameters (assuming that it is gaussian or log gaussian) in a population of patients, and to maximize the likelihood of the observations of all patients. It is a very useful strategy when few data per patients are available, but when we have a lot of patients. The number of parameters to find is multiplied by two (average and standard deviation for each parameter) but the number of data is greatly increased.

This strategy, that we will call "population" parametrization has been initiated in the eighties by software like Nonmem. Recently Marc Lavielle (Popix team) made a series of breakthroughs and designed a new powerfull algorithm, leading to Monolix software.

However population parametrization is very costly. It requires several hundred of thousands of model evaluations, which may be very long.

3.4.2. Results

We address the problem of computation time when the complex model is long to evaluate. In simple cases like reaction diffusion equations in one space dimension, the evaluation of the model may take a few seconds of even a few minutes. In more realistic geometries, the computation time would be even larger and can reach the hour or day. It is therefore impossible to run a SAEM algorith on such models, since it would be much too long. Moreover the underlying algorithm can not be parallelized.

We propose a new iterative approach combining a SAEM algorithm together with a kriging. This strategy appears to be very efficient, since we were able to parametrize a PDE model as fast as a simple ODE model.

We are currently developing the corresponding software.

3.5. Models for the analysis of efficacy data in oncology

3.5.1. Project-team positioning

The development of new drugs for oncology patients faces significant issues with a global attrition rate of 95 percents and only 40 percents of drug approval in phase III after successful phase II. As for meteorology, the analysis through modeling and simulation (MS), of time-course data related to anticancer drugs efficacy and/or toxicity constitutes a rational method for predicting drugs efficacy in patients. This approach, now supported by regulatory agencies such as the FDA, is expected to improve the drug development process and in consequence the treatment of cancer patients. A private company, Pharsight, has nowadays the leader team in the development of such modeling frameworks. In 2009, this team published a model describing tumor size time-course in more than one thousand colorectal cancer patients. This model was used in an MS framework to predict the outcome of a phase III clinical trials based on the analysis of phase II data. From 2009 to 2013, 12 published articles address similar analysis of different therapeutic indications such as lung, prostate, thyroid and renal cancer. A similar modeling activity is also proposed for the analysis of data in preclinical experiments, and in particular, experiments in mice. Animal experiments represent critical stages to decide if a drug molecule should be tested in humans. MS methods are considered as tools to better investigate the mechanisms of drug action and to potentially facilitate the transition towards the clinical phases of the drug development process. Our team has worked in the development of two modeling frameworks with application in both preclinical and clinical oncology. For the preclinical context, we have worked on the development of models focusing on the process of tumor angiogenesis, i.e. the formation of intra-tumoral blood vessels. At the clinical level, we have developed a model to predict tumor size dynamics in patients with low-grade glioma.

At Inria, several project-teams have developed similar efforts. The project-team BANG has a solid experience in the development of age-structured models of the cell cycle and tissue regulation of tumors with clinical applications for chronotherapy. BANG is also currently applying these types of partial differential equation (PDE) models to the study of leukemia through collaboration with the project-team DRACULA. Project-team MC2 has recently shown that the analysis, through a simplified PDE model of tumor growth and treatment response, of 3D imaging, could lead to correct prediction of tumor volume evolution in patients with pulmonary metastasis from thyroid cancer. Regarding specifically the modeling of brain tumors, project-team ASCLEPIOS has brought an important contribution towards personalized medicine in analyzing 3D data information from MRI with a multiscale model that describes the evolution of high grade gliomas in the brain. Their framework relies on the cancer physiopathological model that was mainly developed by Kristin Swanson and her group at the university of Washington.

Outside from Inria, we wish to mention here the work of the group of Florence Hubert in Marseille in the development of models with an interesting compromise between mathematical complexity and data availability. A national ANR project led by the team is expected to support the development of an MS methodology for the analysis of tumor size data in patients with metastases.

3.5.2. Results

Regarding our contribution in preclinical modeling, we have developed a model to analyze the dynamics of tumor progression in nude mice xenografted with HT29 or HCT116 colorectal cancer cells. This model, based on a system of ordinary differential equations (ODEs), integrated the different types of tumor tissues, and in particular, the proliferating, hypoxic and necrotic tissues. Practically, in our experiment, tumor size was periodically measured, and percentages of hypoxic and necrotic tissue were assessed using immunohistochemistry techniques on tumor samples after euthanasia. In the proposed model, the peripheral non-hypoxic tissue proliferates according to a generalized-logistic equation where the maximal tumor size is represented by a variable called "carrying capacity". The ratio of the whole tumor size to the carrying capacity was used to define the hypoxic stress. As this stress increases, non-hypoxic tissue turns hypoxic. Hypoxic tissue does not stop proliferating, but hypoxia constitutes a transient stage before the tissue becomes necrotic. As the tumor grows, the carrying capacity increases owing to the process of angiogenesis (Ribba, Watkin et al. 2011). The model

is shown to correctly predict tumor growth dynamics as well as percentages of necrotic and hypoxic tissues within the tumor.

Regarding our contribution in clinical oncology, we developed an ODE model based on the analysis of mean tumor diameter (MTD) time-course in low-grade glioma patients (Ribba, Kaloshi et al. 2012).

In this model, the tumor is composed of proliferative (P) and non-proliferative quiescent tissue (Q) expressed in millimeters. The proportion of proliferative tissue transitioning into quiescence is constant. The treatment directly eliminates proliferative cells by inducing lethal DNA damage while these cells progress through the cell cycle. The quiescent cells are also affected by the treatment and become damaged quiescent cells (k_{PQ}). Damaged quiescent cells, when re-entering the cell cycle, can repair their DNA and become proliferative once again (transition from Q_P to P) or can die due to unrepaired damages. We modeled the pharmacokinetics of the PCV chemotherapy using a kinetic-pharmacodynamic (K-PD) approach, in which drug concentration is assumed to decay according to an exponential function. In this model, we did not consider the three drugs separately. Rather, we assumed the treatment to be represented as a whole by a unique variable (C), which represents the concentration of a virtual drug encompassing the three chemotherapeutic components of the PCV regimen. We modeled the exact number of treatment cycles administered by setting the value of C to 1 (arbitrary unit) at the initiation of each cycle(T_{Treat}): $C(T = T_{Treat}) = 1$.

The resulting model is as follows:

$$\begin{aligned}
\frac{dC}{dt} &= -KDE \times C \\
\frac{dP}{dt} &= \lambda_P P \left(1 - \frac{P^{\bigstar}}{K}\right) + k_{Q_pP} Q_p - k_{PQ} P - \gamma \times C \times KDE \times P \\
\frac{dQ}{dt} &= k_{PQ} P - \gamma \times C \times KDE \times Q \\
\frac{dQ_p}{dt} &= \gamma \times C \times KDE \times Q - k_{Q_pP} Q_p - \delta_{Q_p} Q_p
\end{aligned} \tag{20}$$

We challenged this model with additional patient data. In particular, MTD time-course information from 24 patients treated with TMZ (subset of the 120 patients from SH) and 25 patients treated with radiotherapy (SH). Note that exactly the same K-PD approach was used to model treatment pharmacokinetic (including for radiotherapy). This choice, though not really realistic was adopted for simplicity reasons: the same model can be indifferently applied to the three different treatment modalities of LGG patients.

3.5.3. Collaborations

François Ducray and Jérôme Honnorat (Pierre Wertheimer Hospital in Lyon)

External support: grant INSERM PhysiCancer 2012 and Inria IPL MONICA

3.6. Stroke

3.6.1. Project team positioning

Stroke is a major public health problem since it represents the second leading cause of death worldwide and the first cause of acquired disability in adults.

Numed is currently starting completely new issues with D. Rousseau (INSA) and his team. We have now at hand a large data base of clinical images. Our aim is to develop model which are able to predict the final size of the dead brain area as a function of the first two clinical data.

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) "naive" 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 [73] in the frame of combustion. It was later used to develop the Kiva code [63] at the Los Alamos National Laboratory, or the Hesione code [68], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [66], 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 [69]. 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 project [0]. 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 [72], and Biot [64] 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 [65]), 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. Intraprenchymal airways distal from generation 7 of the tracheabronchial 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.