



RESEARCH CENTER

FIELD

**Digital Health, Biology and Earth**

Activity Report 2014

# Section Scientific Foundations

Edition: 2015-03-24



## COMPUTATIONAL BIOLOGY

1. ABS Project-Team	5
2. AMIB Project-Team	9
3. BAMBOO Project-Team	19
4. BEAGLE Project-Team	23
5. BIGS Project-Team	27
6. BONSAI Project-Team	31
7. DYLISS Project-Team	32
8. GENSCALE Project-Team	37
9. IBIS Project-Team	38
10. LIFEWARE Team	43
11. MAGNOME Project-Team	46
12. MORPHEME Project-Team	48
13. SERPICO Project-Team	50
14. VIRTUAL PLANTS Project-Team	52

## COMPUTATIONAL NEUROSCIENCE AND MEDECINE

15. ARAMIS Project-Team	54
16. ASCLEPIOS Project-Team	56
17. ATHENA Project-Team	59
18. DEMAR Project-Team	63
19. GALEN Project-Team	66
20. MNEMOSYNE Project-Team	70
21. NEUROMATHCOMP Project-Team	74
22. NEUROSYS Team	77
23. PARIETAL Project-Team	79
24. POPIX Team	83
25. SHACRA Project-Team	84
26. SISTM Team	87
27. VISAGES Project-Team	89

## EARTH, ENVIRONMENTAL AND ENERGY SCIENCES

28. ANGE Project-Team	91
29. CASTOR Project-Team	95
30. CLIME Project-Team	97
31. COFFEE Project-Team	99
32. FLUMINANCE Project-Team	101
33. KALIFFE Project-Team	106
34. LEMON Team	109
35. MAGIQUE-3D Project-Team	114
36. MOISE Project-Team	118
37. POMDAPI Project-Team (section vide)	121
38. SAGE Project-Team	122

39. STEEP Team .....	123
40. TONUS Team .....	126
MODELING AND CONTROL FOR LIFE SCIENCES	
41. BIOCORE Project-Team .....	130
42. CARMEN Team .....	132
43. DRACULA Project-Team .....	134
44. M3DISIM Team .....	137
45. MAMBA Team .....	138
46. MASAIE Project-Team .....	140
47. MODEMIC Project-Team .....	143
48. MYCENAE Project-Team .....	146
49. NUMED Project-Team .....	149
50. REO Project-Team .....	155
51. SISYPHE Project-Team (section vide) .....	158

## 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<sup>0</sup>, 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 [39]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [42]. Current investigations follow two routes. From the experimental perspective [25], 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 [36]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [31].

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<sup>0</sup>, 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 [20], 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 [40], [27]. 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 [32]. 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.

<sup>0</sup>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.

<sup>0</sup>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 [8]. 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 [21]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [41], 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  $C_\alpha$  carbons surrounding a hydrogen bond [24].

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 [35]. 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 [19]. 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 [18], 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 [17], 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 [17]. 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<sup>0</sup>. 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 [23]. 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) [38]. 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 [34], to Morse theory [29] and to analysis of meta-stable states of time series [30] 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.

---

<sup>0</sup>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 [33].

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 [5]. 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.

3D modeling was supported by the Digiteo project JAPARIN-3D. Statistical potentials were supported by CARNAGE and ITSNAPE.

##### 3.1.1. Dynamic programming and complexity

**Participants:** Alain Denise, Yann Ponty, Antoine Soulé.

*Common activity with J. Waldispühl (McGill).*

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 [77], the correction of pyrosequencing errors [71], and the efficient detection of multi-stable RNAs (riboswitches) [72], [73].

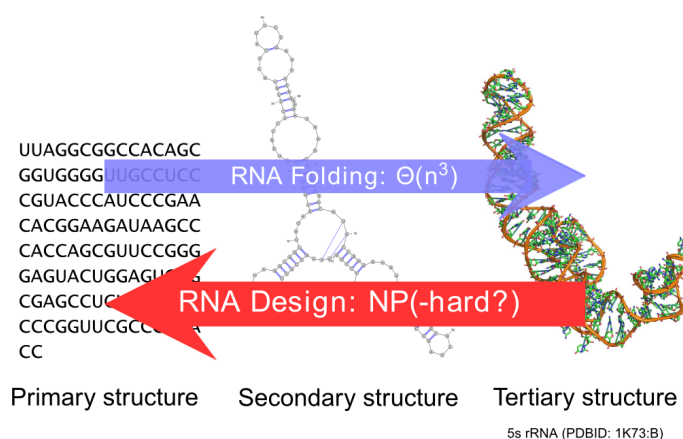


Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

### 3.1.2. RNA design.

**Participants:** Alain Denise, Vincent Le Gallic, Yann Ponty.

*Joint project with 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 [54] 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 [60].

### 3.1.3. Towards 3D modeling of large molecules

**Participants:** Alain Denise, Mélanie Boudard.

*Joint project with D. Barth (Versailles) and J. Cohen (Paris-Sud).*

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 work in both directions.

### 3.1.4. Statistical and robotics-inspired models for structure and dynamics

**Participants:** Julie Bernauer, Rasmus Fonseca.

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 3.1.4 .

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 [76], [51], [64], [63]. 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 [62]. For RNA, progress in this direction has been realized as well. A kinematics-based conformational sampling algorithm, KGS, for loops was recently developed [58], 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 [58], [76] 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 [2], [74] 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 molecules. Rasmus Fonseca, post-doctoral scholar in the project is currently performing an extensive computational and biological validation.

## 3.2. Sequences

**Participants:** Alain Denise, Mireille Régnier, 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. 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. 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 [66]. 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. Combinatorics of motifs

**Participants:** Mireille Régnier, 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.

Molecular interactions often involve specific motifs. One may cite protein-DNA (cis-regulation), protein-protein (docking), RNA-RNA (miRNA, frameshift, circularisation). Motif detection combines an algorithmic search of potential sites and a significance assessment. Assessment significance requires a quantitative criterium. It is generally accepted that the p-value is a reliable tool that outperforms older criteria such as the z-score. AMIB develops a long term research on word combinatorics. 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 word overlaps in a graph [40]. Recursive equations to compute pvalues may be reduced to a traversal of that graph, leading to a linear algorithm. It allows for a derivation of pvalues, decreasing the space and time complexity of the generating function approach or previous probabilistic weighted automata.

In the mean time, continuous sequences of overlapping words, currently named *clumps* or *clusters* turn out to be crucial in random words counting. Notably, they play a fundamental role in the Chen-Stein method of compound Poisson approximation. A first characterization was proposed by Nicodème and al. and this work is currently extended.

This research area is widened by new problems arising from *de novo* genome assembly or re-assembly. For example, unique mappability of short reads strongly depends of the repetition of words. Although the average values for the length have been studied for long under different constraints, their distribution or profile remained unknown until the seminal paper [67] which provides formulae for binary tries. A collaboration has been started with LOB at Ecole Polytechnique to check these formulae on real data, namely Archae genomes (internship of D. Busatto-Gaston). This collaboration has been extended since LOB bought a sequencing machine and a co-advised thesis (Alice Héliou) on circular RNA characterization has just started.

As a third example, one objective is to develop a model of errors, including a statistical model, that takes into account the quality of data for the different sequencing technologies, and their volume. This is the subject of an international collaboration with V. Makeev's lab (IoGene, Moscow) and MAGNOME project-team. Finally, Next Generation Sequencing open the way to the study of structural variants in the genome, as recently described in [48]. Defining a probabilistic model that takes into account main dependencies -such as the GC content- is a task of D. Iakovishina's thesis, to be defended in 2015, in a collaboration with V. Boeva (Curie Institute).

### 3.2.2. *Random generation*

**Participants:** Alain Denise, 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 [55]: 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, we have done significant and original progress in this area recently [68], [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 [66].

Besides, our work on random generation is also applied in a different fields, namely software testing and model-checking, in a continuing collaboration with the Fortesse group at LRI [53], [65].

## 3.3. Geometry and machine learning for 3D interaction prediction

**Participants:** Julie Bernauer, Jean-Marc Steyaert, Christine Froidevaux, Adrien Guilhot-Gaudeffroy, 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 [47], [57] and efficient code are essential [61].

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. Modelling is then raised to the cell level by studying large protein interaction networks and also the dynamics of molecular pathways.

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. We also contributed to the CAPRI experiment organized by leading member of an international community we have been involved in for some time [56]. At Inria, our interest for structural biology is shared by the ABS project-team. A work by D. Ritchie in the ORPAILLEUR project-team (see [44] led to a joint publication with T. Bourquard and J. Azé. 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.

### 3.3.1. Combinatorial models for the structure of proteins

Protein structure prediction has been and still is extensively studied. Computational approaches have shown interesting results for globular proteins but transmembrane proteins remain a difficult case.

Transmembrane beta-barrel proteins (TMB) account for 20 to 30% of identified proteins in a genome but, due to difficulties with standard experimental techniques, they are only 2% of the RCSB Protein Data Bank. As TMB perform many vital functions, the prediction of their structure is a challenge for life sciences, while the small number of known structures prohibits knowledge-based methods for structure prediction.

As barrel proteins are strongly structured objects, model based methodologies are an interesting alternative to these conventional methods. Jérôme Waldispühl's thesis at LIX had opened this track for the common case where a protein folds respecting the order of the sequence, leaving a structure where each strand is bound to the preceding and succeeding ones. The matching constraints were expressed by a grammatical model, for which relatively simple dynamic programming schemes exist.

However, more sophisticated schemes are required when the arrangements of the strands along the barrel do not follow their order in the sequence, as it is the case for *Greek key* or *Jelly roll* motifs. The prediction algorithm may then be driven by a permutation on the order of the bonded strands. In his thesis [75], Van Du Tran developed a methodology for compiling a given permutation into a dynamic programming scheme that may predict the folding of sequences into the corresponding TMB secondary structure. Polynomial complexity upper bounds follow from the calculated DP scheme. Through tree decompositions of the graph that expresses constraints between strands in the barrel, better schemes were investigated in [75].

The efficiently obtained 3D structures provide a good model for further 3D and interaction analyses.

### 3.3.2. 3D interaction prediction

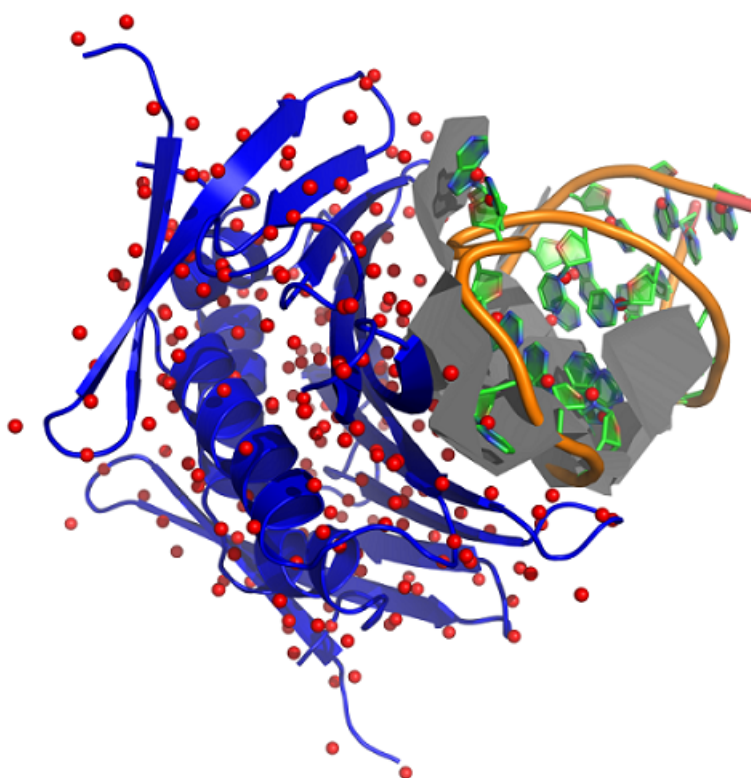
To better model complexes, various aspects of the scoring problem for protein-protein docking need being addressed [56]. It is also of great interest to introduce a hierarchical analysis of the original complex three-dimensional structures used for learning, obtained by clustering.

A protein-protein docking procedure traditionally consists in two successive tasks: a search algorithm generates a large number of candidate solutions, and then a scoring function is used to rank them in order to extract a native-like conformation. We demonstrated that, using Voronoi constructions and a defined set of parameters, we could optimize an accurate scoring function and interaction detection [46]. We also focused on developing other geometric constructions for that purpose: being related to the Voronoi construction, the Laguerre tessellation was expected to better represent the physico-chemical properties of the partners. It also allows a fast computation without losing the intrinsic properties of the biological objects. In [49], we compare both constructions. We also worked on introducing a hierarchical analysis of the original complex three-dimensional structures used for learning, obtained by clustering. Using this clustering model, in combination with a strong emphasis on the design of efficient complex filters collaborative filtering, we can optimize the scoring functions and get more accurate solutions [50].

These techniques have been extended to the analysis of protein-nucleic acid complexes : developments and tests are performed by A. Guilhot (See figure 2 ) in his PhD thesis.

## 3.4. Data Integration

**Participants:** Christine Froidevaux, Alain Denise, Sarah Cohen-Boulakia, Bryan Brancotte, Jiuqiang Chen.



*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.*

Faced with the inherent features of biological and biomedical data, researchers from the database and artificial intelligence communities have joined together to form a community dedicated to the study of the specific problems posed by integrating life sciences data. With the deluge of new sequenced genome sequences and the amount of data produced by high-throughput approaches, the need to cross and compare massive and heterogeneous data is more important than ever to improve functional annotation and design biological networks. Challenges are numerous. One may cite the need to provide support to scientists to perform and share complex and reproducible complex biological analyses. A special attention is paid to the more specific domain of scientific workflows management and ranking biological data. One aims at exploring the relationships between those two domains, from the investigation of various specific problems posed by ranking scientific workflows to the problem of considering consensus workflows.

### 3.4.1. Designing and Comparing Scientific workflows

**Participants:** Sarah Cohen-Boulakia, Christine Froidevaux, Jiuqiang Chen.

Scientific workflows management systems are increasingly used to specify and manage bioinformatics experiments. Their programming model appeals to bioinformaticians, who use them to easily specify complex data processing pipelines. Such a model is underpinned by a graph structure, where nodes represent bioinformatics tasks and links represent the dataflow. As underlined both in a study and a review of existing approaches, the complexity of such graph structures is increasing over time, making them more difficult to share and reuse.

One of the major current challenges is thus to provide means to reduce the structural complexity of workflows while ensuring that any structural transformation will not have any impact on the executions of the transformed workflows, that is, preserving *provenance*.

### 3.4.2. Ranking biological data

**Participants:** Alain Denise, Sarah Cohen-Boulakia, Bryan Brancotte.

We are addressing the increase of the number of resources available. The BIOGUIDE project aim at helping user navigation in the maze of available biological sources. More recently, a second problem was tackled: the number of answers returned by even one single queried biological resource may be too large for the user to deal with. We have provided solutions for ranking biological data. The main difficulty lies in considering various ranking criteria (recent data first, popular data first, curated data first...). Many approaches combine ranking criteria to design a ranking function, possibly leading to arbitrary choices made in the way of combining the ranking criteria. Instead, in collaboration with the University of Montreal, we have proposed to follow a *median ranking approach* named BIOCONSERT (for generating Biological Consensus ranking with ties): considering as many rankings as they are ranking criteria for the same data set, and providing a consensus ranking that minimizes the disagreements between the input rankings. We have shown the benefit of using median ranking in several biological settings.

Additionally, in a close collaboration with the Institut Curie, we have also developed the GENEVALORIZATION tool that ranks a list of genes of interest given as input with respect to a set of keywords representing the context of study. Here the single ranking criterion considered for each gene is the number of publications in PubMed co-citing the gene name and the keywords. The tool is able to make use of the MeSH taxonomy when considering the keywords and the dictionary of gene names and aliases for the gene names.

## 3.5. Systems Biology

**Participants:** Patrick Amar, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Loic Paulevé, Sabine Pérès, Jean-Marc Steyaert, Erwan Bigan, Adrien Rougny.

Systems Biology involves the systematic study of complex interactions in biological systems using an integrative approach. The goal is to find new emergent properties that may arise from the systemic view in order to understand the wide variety of processes that happen in a biological system. Systems Biology activity can be seen as a cycle composed of theory, computational modelling to propose a hypothesis about a biological process, experimental validation, and use of the experimental results to refine or invalidate the computational model (or even the whole theory). During the past five years, new questions and research domains have been identified, and some members of the team have reoriented a part of their activities on these questions.

Three main types of problems have been studied: metabolic networks, signaling networks and more recently synthetic biology. Networks - have become popular since many crucial problems, coming from biology, medicine, pharmacology, are nowadays stated in these terms: a great number of them are issued from the cancer phenomenon and the will to enhance our understanding in order to propose more efficient therapeutic issues. Metabolism has received the major attention since it concerns a large variety of topics and several methods that have been proposed. Depending on the nature of the biological problem, several methods can be used : discrete deterministic, stochastic, combinatorial, up to continuous differential. Also, the recent rise of synthetic biology proposes similar challenges aiming at improving the production of energy by means of biological systems or at getting more efficient medicament treatments, for instance.

### 3.5.1. *Topological analysis of metabolic networks*

**Participant:** Sabine Pérès.

Elementary flux mode analysis is a powerful tool for the theoretical study of simple metabolic networks. However, when the networks are complex, the determination of elementary flux modes leads to a combinatorial explosion of their number which prevents from drawing simple conclusions from their analysis. Since the concept of elementary flux mode analysis was introduced in 1994, there has been an important and ongoing effort to develop more efficient algorithms. However, these methods share a common bottleneck: they enumerate all the elementary flux modes which make the computation impossible when the metabolic network is large and only few works try to search only elementary flux mode with specific properties. We have shown that enumerating all the elementary flux modes is not necessary in many cases and it is possible to directly query the network instead with an appropriate tool. For ensuring a good query time, we have relied on a state of the art SAT solver, working on a propositional encoding of elementary flux mode, and enriched with a simple SMT-like solver ensuring elementary flux mode consistency with stoichiometric constraints. We have illustrated our new framework by providing experimental evidences of almost immediate answer times on a non trivial metabolic network [45], [70].

### 3.5.2. *Signaling networks*

**Participants:** Sarah Cohen-Boulakia, Christine Froidevaux, Adrien Rougny.

Signaling pathways involving G protein-coupled receptors (GPCR) are excellent targets in pharmacogenomics research. Large amounts of experiments are available in this context while globally interpreting all the experimental data remains a very challenging task for biologists. Our goal is to help the understanding of signaling pathways involving (GPCR) and to provide means to semi-automatically construct the signaling networks.

We have introduced a logic-based method to infer molecular networks and show how it allows inferring signaling networks from the design of a knowledge base. Provenance of inferred data has been carefully collected, allowing quality evaluation. Our method (i) takes into account various kinds of biological experiments and their origin; (ii) mimics the scientist's reasoning within a first-order logic setting; (iii) specifies precisely the kind of interaction between the molecules; (iv) provides the user with the provenance of each interaction; (v) automatically builds and draws the inferred network [43].

Observe that a logic-based formalisation is used as in some works carried out in INRIA team DYLISS. AMIB aim is different, as the design of the network lies on a knowledge-based system describing experimental facts and ontological relationships on background knowledge, together with a set of generic and expressive rules, that mimick the expert's reasoning.

This is a collaboration with A. Poupon (INRA-BIOS, Tours) that was supported by an INRA-INRIA starting grant in 2011-2012.

### 3.5.3. *Modelling and Simulation*

**Participants:** Patrick Amar, Sarah Cohen-Boulakia, Loic Paulevé, Jean-Marc Steyaert, Erwan Bigan.



A great number of methods have been proposed for the study of the behavior of large biological systems. The first one is based on a discrete and direct simulation of the various interactions between the reactants using an entity-centered approach; the second one implements a very efficient variant of the Gillespie stochastic algorithm that can be mixed with the entity-centered method to get the best of both worlds; the third one uses differential equations automatically generated from the set of reactions defining the network.

These three methods have been implemented in an integrated tool, the HSIM system [41]. It mimics the interactions of biomolecules in an environment modelling the membranes and compartments found in real cells. It has been applied to the modelling of the circadian clock of the cyanobacterium, and we have shown pertinent results regarding the spontaneous appearance of oscillations and the factors governing their period [42].

#### 3.5.3.1. Synthetic biology

Synthetic biology begins to be a very popular domain of research. Genetic engineering is a good example of synthetic biology, organisms are artificially modified to boost the production of compounds that might be used in the medical or industrial domains. We have been focused on using synthetic biology for medical diagnostic purposes. In a collaboration with the SYSDIAGLab (UMR 3145) at Montpellier, P. Amar participates at the COMPUBIOTIC project. The goal is to design, test and build an artificial embedded biological nano-computer in order to detect the biological markers of some human pathologies (colorectal cancer, diabetic nephropathy, etc.). This nano-computer is a small vesicle containing specific enzymes and membrane receptors. These components are chosen in a way that their interactions can sense and report the presence in the environment of molecules involved in the human pathologies targeted. We plan to design a dedicated software suite to help the design and validation of this artificial nano-computer. HSIM is used to help the design and to test qualitatively and quantitatively this "biological computer" before *in vitro*.

#### 3.5.3.2. Evaluating metabolic networks

It is now well established in the medical world that the metabolism of organs depends crucially of the way they consume oxygen, glucose and the various metabolites that allow them to grow and duplicate. A particular variety of cells, tumour cells, is of major interest. In collaboration with L. Schwartz (AP-HP) and biologists from INSERM-INRA Clermont-Theix we have started a project aiming at identifying the important points in the metabolic machinery that command the changes in behaviour. The main difficulties come from the fact that biologists have listed dozens of concurrent cycles that can be activated alternatively or simultaneously, and that the dynamic characteristics of the chemical reactions are not known accurately.

Given the set of biochemical reactions that describe a metabolic function (e.g. glycolysis, phospholipids' synthesis, etc.) we translate them into a set of o.d.e's whose general form is most often of the Michaelis-Menten type but whose coefficients are usually very badly determined. The challenge is therefore to extract information as to the system's behavior while making reasonable assumptions on the ranges of values of the parameters. It is sometimes possible to prove mathematically the global stability, but it is also possible to establish it locally in large subdomains by means of simulations. Our program Mpas (Metabolic Pathway Analyser Software) renders the translation in terms of a systems of o.d.e's automatic, leading to easy, almost automatic simulations. Furthermore we have developed a method of systematic analysis of the systems in order to characterize those reactants which determine the possible behaviors: usually they are enzymes whose high or low concentrations force the activation of one of the possible branches of the metabolic pathways. A first set of situations has been validated with a research INSERM-INRA team based in Clermont-Ferrand. In her PhD thesis, defended in 2011, M. Behzadi proved mathematically the decisive influence of the enzyme PEMT on the Choline/Ethylamine cycles.

#### 3.5.3.3. Comparison of Metabolic Networks

We study the interest of *fungi* for biomass transformation. Cellulose, hemicellulose and lignin are the main components of plant biomass. Their transformation represent a key energy challenges of the 21st century and should eventually allow the production of high value new compounds, such as wood or liquid biofuels (gas or bioethanol). Among the boring organisms, two groups of *fungi* differ in how they destroy the wood compounds. Analysing new *fungi* genomes can allow the discover of new species of high interest for

bio-transformation. For a better understanding of how the fungal enzymes facilitates degradation of plant biomass, we conduct a large-scale analysis of the metabolism of fungi. Machine learning approaches such like hierarchical rules prediction are being studied to find new enzymes allowing the transformation of biomass. The KEGG database <http://www.genome.jp/kegg/> contains pathways related to fungi and other species. By analysing these known pathways with rules mining approaches, we aim to predict new enzymes activities.

## BAMBOO Project-Team

### 3. Research Program

#### 3.1. Symbiosis

The study we propose to do on symbiosis decomposes into four main parts - (1) genetic dialog, (2) metabolic dialog, (3) symbiotic dialog and genome evolution, and (4) symbiotic dynamics - that are however strongly interrelated, and the study of such interrelations will represent an important part of our work. Another biological objective, larger and which we hope within the ERC project SISYPHE just to sketch for a longer term investigation, will aim at getting at a better grasp of species identity and of a number of identity-related concepts. We now briefly indicate the main points that have started been investigated or should be investigated in the next five years.

##### Genetic dialog

We plan to study the genetic dialog at the regulation level between symbiont and host by addressing the following mathematical and algorithmic issues:

1. model and identify all small RNAs from the bacterium and the host which may be involved in the genetic dialog between the two, and model/identify the targets of such small RNAs;
2. infer selected parts of the regulatory network of both symbiont and host (this will enable to treat the next point) using all available information;
3. explore at both the computational and experimental levels the complementarity of the two networks, and revisit at a network level the question of a regulatory response of the symbiont to its host's demand;
4. compare the complementarities observed between pairs of networks (the host's and the symbiont's); such complementarities will presumably vary with the different types of host-symbiont relationships considered, and of course with the information the networks model (structural or dynamic); Along the way, it may become important at some point to address also the issue of transposable elements (abbreviated into TEs, that are genes which can jump spontaneously from one site to another in a genome following or not a duplication event). It is increasingly believed that TEs play a role in the regulation of the expression of the genes in eukaryotic genomes. The same role in symbionts, and in the host-symbiont dialog has been less or not explored. This requires to address the following additional task:
5. accurately and systematically detect all transposable elements (*i.e.* genes which can jump spontaneously from one site to another in a genome following or not a duplication event) and assess their implication in their own regulation and that of their host genome (the new sequencing technologies should facilitate this task as well as other data expression analyses, if we are able to master the computational problem of analysing the flow of data they generate: fragment indexing, mapping and assembly);
6. where possible, obtain data enabling to infer the PPI (Protein-Protein Interaction) for hosts and symbionts, and at the host-symbiont interface and analyse the PPI networks obtained and how they interact.

Initial algorithmic and statistical approaches for the first two items above are under way and are sustained by a well-established expertise of the team on sequence and microarray bioinformatic analysis. Both problems are however notoriously hard because of the high level of missing data and noise, and of our relative lack of knowledge of what could be the key elements of genetic regulation, such as small and micro RNAs.

We also plan to establish the complete repertoire of transcription factors of the interacting partners (with possible exchanges between them) at both the computational and experimental levels. Comparative biology (search by sequence homology of known regulators), 3D-structural modelling of putative domains interacting with the DNA molecule, regulatory domains conserved in the upstream region of coding DNA are among classical and routinely used methods to search for putative regulatory proteins and elements in the genomes. Experimentally, the BiaCore (using the surface plasmon resonance principle) and ChIP-Seq (using chromatin precipitation coupled with high-throughput sequencing from Solexa) techniques offer powerful tools to capture all the protein-DNA interactions corresponding to a specific putative regulator. However, these techniques have not been evaluated in the context of interacting partners making this task an interesting challenge.

### **Metabolic dialog**

Our main plan for this part, where we have already many results, some obtained this last year, is to:

1. continue with and improve our work on reconstructing the metabolic networks of organisms with sequenced genomes, taking in particular care to cover as much as possible the different types of hosts and symbionts in interaction;
2. refine the network reconstructions by using flux balance analysis which will in turn require addressing the next item;
3. improve our capacity to efficiently compute fluxes and do flux balance analysis; current algorithms can handle only relatively small networks;
4. analyse and compare the networks in terms of their general structural, quantitative and dynamic characteristics;
5. develop models and algorithms to compare different types of metabolic interfaces which will imply being able, by a joint computational and experimental approach, to determine what is transported across interacting metabolisms;
6. define what would be a good null hypothesis to test the statistical significance, and therefore possible biological relevance of the characteristics observed when analysing or comparing (random network problem, a mostly open issue despite the various models available);
7. use the results from item 5, that is indications on the precursors of a bacterial metabolism that are key players in the dialog with the metabolism of the host, to revisit the genetic regulation dialog between symbiont and host.

Computational results from the last item will be complemented with experiments to help understand what is transported from the host to the symbiont and how what is transported may be related with the genetic dialog between the two organisms (items 5 and 6).

Great care will also be taken in all cases (metabolism- or regulation-only, or both together) to consider the situations, rather common, where more than two partners are involved in a symbiosis, that is when there are secondary symbionts of a same host.

The first five items above have started being computationally explored by our team, as has the last item including experimentally. Some algorithmic proofs-of-concept, notably as concerns structural, flux, precursor and chemical organisation studies (see some of the publications of the last year and this one), have been established but much more work is necessary. The main difficulties with items 3 and 4 are of two sorts. The first one is a modelling issue: what are the best models for analysing and comparing two or more networks? This will greatly depend on the biological question put, whether evolutionary or functional, structural or physiologic, besides being a choice that should be motivated by the extent and quality of the data available. The second sort of difficulty, which also applies to other items notably (item 2), is computational. Most of the problems related with analysing and specially comparing are known to be hard but many issues remain open. The question of a good random model (item 6) is also largely open.

### **Symbiotic dialog and genome evolution**

Genomes are not static. Genes may get duplicated, sometimes the duplication affects the whole genome, or genes can transpose, while whole genomic segments can be reversed or deleted. Deletions are indeed one of the most common events observed for some symbionts. Genetic material may also be transferred across sub-species or species (lateral transfer), thus leading to the insertion of new elements in a genome. Finally, parts of a genome may be amplified through, for instance, slippage during DNA replication resulting in the multiplication of the copies of a repeat that appear tandemly arrayed along a genome. Tandem repeats, and other types of short or long repetitions are also believed to play a role in the generation of new genomic rearrangements although whether they are always the cause or consequence of the genome break and gene order change remains a disputed issue.

Work on this part will involve the following items:

1. extend the theoretical work done in the past years (rearrangement distance, rearrangement scenarios enumeration) to deal with different types of rearrangements and explore various types of biological constraints;
2. develop good random models (a largely open question despite some initial work in the area) for rearrangement distances and scenarios under a certain model, i.e. type of rearrangement operation(s) and of constraint(s), to assess whether the distances / scenarios observed have statistically notable characteristics;
3. extensively use the method(s) developed to investigate the rearrangement histories for the families of symbionts whose genomes have been sequenced and sufficiently annotated;
4. investigate the correlation of such histories with the repeats content and distribution along the genomes;
5. use the results of the above analyses together with a natural selection criterion to revisit the optimality model of rearrangement dynamics;
6. extend such model to deal with eukaryotic (multi-chromosomal) genomes;
7. at the interface host-symbiont, investigate the relation between the rearrangement histories in hosts and symbionts and the various types of symbiotic relationships observed in nature;
8. map such histories and their relation with the genetic and metabolic networks of hosts and symbionts, separately and at the interface;
9. develop methods to identify and quantify rearrangement events from NGS data.

### **Symbiotic dynamics**

In order to understand the evolutionary consequences of symbiotic relations and their long term trajectories, one should be able to assess how tight is the association between symbionts and their hosts.

The main questions we would like to address are:

1. how often are symbionts horizontally transferred among branches of the host phylogenetic tree?
2. how long do parasites persist inside their host following the invasion of a new lineage?
3. what processes underlie this dynamic gain/loss equilibrium?

Mathematically, these questions have been traditionally addressed by co-phylogenetic methods, that is by comparing the evolutionary histories of hosts and parasites as represented in phylogenetic trees.

Currently available co-phylogenetic algorithms present various types of limitations as suggested in recent surveys. This may seriously compromise their interpretation with a view to understanding the evolutionary dynamics of parasites in communities. A few examples of limitations are the (often wrong) assumption made that the same rates of loss and gain of parasite infection apply for every host taxonomic group, and the fact that the possibility of multi-infections is not considered. In the latter case, exchange of genetic material between different parasites of a same host could further scramble the co-evolutionary signal. We therefore plan to:

1. better formalise the problem and the different simplifications that could be made, or inversely, should be avoided in the co-phylogeny studies; examples of the latter are the possibility of multi-infections, differential rate of loss and gain of infection depending on the host taxonomic group and geographic distance between hosts, etc., and propose better co-phylogenetic algorithms;
2. elaborate series of simulated data that will enable to (i) get a better grasp of the effect of the different parameters of the problem and, more practically, (ii) evaluate the performance of the method(s) that exist or are proposed (see next item);
3. apply the new methods to address the three questions above.

### **3.2. Intracellular interactions**

The interactions of a symbiont with others sharing a same host, or with a symbiont and the cell of its host in the case of endosymbionts (organism that lives within the body or cells of another) are special, perhaps more complex cases of intracellular interactions that may concern different types of genetic elements, from organelles to whole chromosomes. The spatial arrangement of those genetic elements inside the nucleus of a cell is believed to be important both for gene expression and exchanges of genetic material between chromosomes. This question goes beyond the symbiosis one and has been investigated in the team in the last few years. Work on this will continue in future and concern developing algorithmic and statistical methods to analyse the interaction data that is starting to become available, in particular using NGS methods, in order to arrive at a better understanding of transcription, regulation both classical and epigenetic (inherited changes in phenotype or gene expression caused by mechanisms other than changes in the underlying DNA sequence), alternative splicing and trans-splicing phenomena, as well as study the possible interactions between an eukaryotic cell and its organelles or other cytoplasmic structures.

## 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 [35]. 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 [39]. 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 [45]. 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 [41]. 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 [52]).

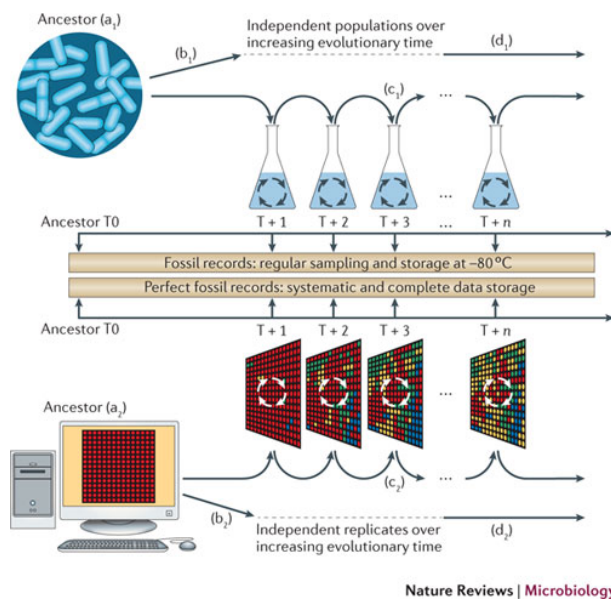


Figure 1. Parallel between experimental evolution and artificial evolution



As a consequence, this model can be used to study how evolutionary pressures like the ones for robustness or evolvability can shape genome structure [53], [50], [51], [60]. 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 [53], [50], [61]. 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 [44]. 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 [37], [36], [38]. 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 [36]. 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 [53], and no difference in lifestyles and environment was needed to explain the complexity of the gene regulatory network [36]. 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, we have started to study how to improve the performance of (R-)Aevol. Thanks to the ADT Aevol, one and a half full time engineers will work to improve 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 Tesla). 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.

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 Adiva and experimental evolution [54], [48], 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 phylogenetic models of punctual mutations on sequences [46], which enable the reconstruction of small parts of ancestral sequences, individual genes for example [55]. But models of whole genome evolution, taking into account large scale events like duplications, insertions, deletions, lateral transfer, rearrangements are just being developed [63], [42]. Integrative phylogenetic models, considering both nucleotide substitutions and genome architectures, like Aevol does, are still missing.

Partial models lead to evolutionary hypotheses on the birth and death of genes [43], on the rearrangements due to duplications [33], [62], on the reasons of variation of genome size [49], [56]. Most of these hypotheses are difficult to test due to the difficulty of *in vivo* evolutionary experiments.

To this aim, we develop evolutionary models for reconstructing 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 [57], methods for reconstructing the organization of ancestral genomes [58], [40], [59], or for detecting lateral gene transfer events [32], [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 will necessitate 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 put into light the relations between different kinds of mutations, and enable the comparison with artificial experiments from 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.

## BIGS Project-Team

### 3. Research Program

#### 3.1. Online data analysis

*Participants: J.-M. Monnez, P. Vallois.* Generally speaking, there exists an overwhelming amount of articles dealing with the analysis of high dimensional data. Indeed, this is one of the major challenges in statistics today, motivated by internet or biostatistics applications. Within this global picture, the problem of classification or dimension reduction of online data can be traced back at least to a seminal paper by Mac Queen [53], in which the  $k$ -means algorithm is introduced. This popular algorithm, constructed for classification purposes, consists in a stepwise updating of the centers of some classes according to a stream of data entering into the system. The literature on the topic has been growing then rapidly since the beginning of the 90's.

Our point of view on the topic relies on the so-called *French data analysis school*, and more specifically on Factorial Analysis tools. In this context, it was then rapidly seen that stochastic approximation was an essential tool (see Lebart's paper [50]), 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 [56], [54], [55], in which many aspects of convergences of online processes are analyzed thanks to the stochastic approximation techniques.

#### 3.2. Local regression techniques

*Participants: S. Ferrigno, A. Muller-Gueudin.* In the context where a response variable  $Y$  is to be related to a set of regressors  $X$ , one of the general goals of Statistics is to provide the end user with a model which turns out to be useful in predicting  $Y$  for various values of  $X$ . Except for the simplest situations, the determination of a good model involves many steps. For example, for the task of predicting the value of  $Y$  as a function of the covariate  $X$ , statisticians have elaborated models such as the regression model with random regressors:

$$Y = g(X, \theta) + \sigma(X)\epsilon.$$

Many assumptions must be made to reach it as a possible model. Some require much thinking, as for example, those related to the functional form of  $g(\cdot, \theta)$ . Some are made more casually, as often those related to the functional form of  $\sigma(\cdot)$  or those concerning the distribution of the random error term  $\epsilon$ . Finally, some assumptions are made for commodity. Thus the need for methods that can assess if a model is concordant with the data it is supposed to adjust. The methods fall under the banner of goodness of fit tests. Most existing tests are *directional*, in the sense that they can detect departures from only one or a few aspects of a null model. For example, many tests have been proposed in the literature to assess the validity of an entertained structural part  $g(\cdot, \theta)$ . Some authors have also proposed tests about the variance term  $\sigma(\cdot)$  (cf. [51]). Procedures testing the normality of the  $\epsilon_i$  are given, but for other assumptions much less work has been done. Therefore the need of a global test which can evaluate the validity of a global structure emerges quite naturally.

With these preliminaries in mind, let us observe that one quantity which embodies all the information about the joint behavior of  $(X, Y)$  is the cumulative conditional distribution function, defined by

$$F(y|x) = P(Y \leq y | X = x).$$

The (nonparametric) estimation of this function is thus of primary importance. To this aim, notice that modern estimators are usually based on the local polynomial approach, which has been recognized as superior to classical estimates based on the Nadaraya-Watson approach, and are as good as the recent versions based on spline and other methods. In some recent works [41], [42], we address the following questions:

- Construction of a global test by means of Cramér-von Mises statistic.
- Optimal bandwidth of the kernel used for approximation purposes.

We also obtain sharp estimates on the conditional distribution function in [43].

### 3.3. Stochastic modeling for complex and biological systems

*Participants:* R. Azais, T. Bastogne, C. Lacaux, A. Muller-Gueudin, S. Tindel, P. Vallois, S. Wantz-Mézières

In most biological contexts, mathematics turn out to be useful in producing accurate models with dual objectives: they should be simple enough and meaningful for the biologist on the one hand, and they should provide some insight on the biological phenomenon at stake on the other hand. We have focused on this kind of issue in various contexts that we shall summarize below.

*Photodynamic Therapy:* Photodynamic therapy induces a huge demand of interconnected mathematical systems, among which we have studied recently the following ones:

- The tumor growth model is of crucial importance in order to understand the behavior of the whole therapy. We have considered the tumor growth as a stochastic equation, for which we have handled the problem uncertainties on the measure times [27] as well as mixed effects for parameter estimation.
- Another important aspect to quantify for photodynamic therapy calibration is the response to radiotherapy treatments. There are several valid mathematical ways to describe this process, among which we distinguish the so-called hit model. This model assumes that whenever a group of sensitive targets (chromosomes, membrane) in the cell are reached by a sufficient number of radiations, then the cell is inactivated and dies. We have elaborated on this scheme in order to take into account two additional facts: (i) The reduction of the cell situation to a two-state model might be an oversimplification. (ii) Several doses of radiations are inoculated as time passes. These observations have lead us to introduce a new model based on multi-state Markov chains arguments (Keinj & al, 2012), in which cell proliferation can be incorporated.

*Bacteriophage therapy:* Let us mention a starting collaboration between BIGS and the Genetics and Microbiology department at the Universitat Autònoma de Barcelona, on the modeling of bacteriophage therapies. The main objective here is to describe how a certain family of benign viruses is able to weaken a bacterium induced disease, which naturally leads to the introduction of a noisy predator-prey system of equations. It should be mentioned that some similar problems have been treated (in a rather informal way, invoking a linearization procedure) by Carletti in [34]. These tools cannot be applied directly to our system, and our methods are based on concentration and large deviations techniques (on which we already had an expertise [57], [60]) in order to combine convergence to equilibrium for the deterministic system and deviations of the stochastic system. Notice that A. Muller-Gueudin is also working with A. Debussche and O. Radulescu on a related topic [37], namely the convergence of a model of cellular biochemical reactions.

*Gaussian signals:* 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 differential equations driven by a fractional Brownian motion (fBm) and other Gaussian processes are now well understood, mainly thanks to the so-called *rough paths* tools [52], but also invoking the Russo-Vallois integration techniques [59]. The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in [38].

Fractional fields are very often used to model irregular phenomena which exhibit a scale invariance property, fractional Brownian motion being the historical fractional model. Nevertheless, its isotropy property is a serious drawback for instance in hydrology or in medicine (see [33]). Moreover, the fractional Brownian motion cannot be used to model some phenomena for which the regularity varies with time. Hence, many generalizations (Gaussian or not) of this model have been recently proposed, see for instance [28] for some Gaussian locally self-similar fields, [46] for some non-Gaussian models, [31] for anisotropic models.

Our team has thus contributed [36], [47], [46], [48], [58] and still contributes [30], [32], [31], [49], [44] to this theoretical study: Hölder continuity, fractal dimensions, existence and uniqueness results for differential equations, study of the laws to quote a few examples. As we shall see below, this line of investigation also has some impact in terms of applications: we shall discuss how we plan to apply our results to osteoporosis on the one hand and to fluctuations within protein molecules on the other hand.

### 3.4. Parameter identifiability and estimation

*Participants: R. Azais, T. Bastogne, S. Tindel, P. Vallois, S. Wantz-Mézières*

When one desires to confront theoretical probabilistic models with real data, statistical tools are obviously crucial. We have focused on two of them: parameter identifiability and parameter estimation.

Parameter identifiability [62] deals with the possibility to give a unique value to each parameter of a mathematical model structure in inverse problems. There are many methods for testing models for identifiability: Laplace transform, similarity transform, Taylor series, local state isomorphism or elimination theory. Most of the current approaches are devoted to *a priori* identifiability and are based on algebraic techniques. We are particularly concerned with *a posteriori* identifiability, *i.e.*, after experiments or in a constrained experimental framework and the link with experimental design techniques. Our approach is based on statistical techniques through the use of variance-based methods. These techniques are strongly connected with global sensitivity approaches and Monte Carlo methods.

The parameter estimation for a family of probability laws has a very long story in statistics, and we refer to [29] for an elegant overview of the topic. Moving to the references more closely related to our specific projects, let us recall first that the mathematical description of photodynamic therapy can be split up into three parametric models : the uptake model (pharmacokinetics of the photosensitizing drug into cancer cells), the photoreaction model and the tumor growth model. Several papers have been reported for the application of system identification techniques to pharmacokinetics modeling problems. But two issues were ignored in these previous works: presence of timing noise and identification from longitudinal data. In [27], 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 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. In the recent past, we have used this tool for tumor growth modeling. According to what we know so far, no parameter estimation study has been reported about the photoreaction model in photodynamic therapy. A photoreaction model, composed of six stochastic differential equations, is proposed in [39]. The main open problem is to access to data. We currently build on an experimental platform which aims at overcoming this technical issue. Moreover, an identifiability study coupled to a global sensitivity analysis of the photoreaction model are currently in progress. Tumor growth is generally described by population dynamics models or by cell cycle models. Faced with this wide variety of descriptions, one of the main open problems is to identify the suitable model structure. As mentioned above, we currently investigate alternative representations based on branching processes and Markov chains, with a model selection procedure in mind.

A few words should be said about the existing literature on statistical inference for diffusion or related processes, a topic which will be at the heart of three of our projects (namely photodynamic and bacteriophage therapies, as well as fluctuations within molecules). The monograph [45] is a good reference on the basic estimation techniques for diffusion processes. The problem of estimating diffusions observed at discrete times,

of crucial importance for applications, has been addressed mainly since the mid 90s. The maximum likelihood techniques, which are also classical for parameter estimation, are well represented by the contributions [40].

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 [26], [35]. On the other hand, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. A good reference on the question is [61], dealing with some very particular families of equations, which do not cover the cases of interest for us.

## **BONSAI Project-Team**

### **3. Research Program**

#### **3.1. Combinatorial discrete models and algorithms**

Our research is driven by biological questions. At the same time, we have in mind to develop well-founded models and efficient algorithms. Biological macromolecules are naturally modelled by various types of discrete structures: String, trees, and graphs. String algorithms is an established research subject of the team. We have been working on spaced seed techniques for several years. Members of the team also have a strong expertise in text indexing and compressed index data structures, such as BWT. Such methods are widely-used for the analysis of biological sequences because they allow a data set to be stored and queried efficiently. Ordered trees and graphs naturally arise when dealing with structures of molecules, such as RNAs or non-ribosomal peptides. The underlying questions are: How to compare molecules at structural level, how to search for structural patterns ? String, trees and graphs are also useful to study genomic rearrangements: Neighborhoods of genes can be modelled by oriented graphs, genomes as permutations, strings or trees. High-performance computing is another tool that we use to achieve our goals.

#### **3.2. Discrete statistics and probability**

At a lower level, our work relies on a basic background on discrete statistics and probability. When dealing with large input data sets, it is essential to be able to discriminate between noisy features observed by chance from those that are biologically relevant. The aim here is to introduce a probabilistic model and to use sound statistical methods to assess the significance of some observations about these data. Examples of such observations are the length of a repeated region, the number of occurrences of a motif (DNA or RNA), the free energy of a conserved RNA secondary structure, etc. Probabilistic models are also used to describe genome evolution. In this context, Bayesian models and MCMC sampling allow to approximate probability distributions over free parameters and to describe biologically relevant models.

## 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) [5].

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** [45], [53] 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 [46], and proposed robust corrections [4]. 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



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.

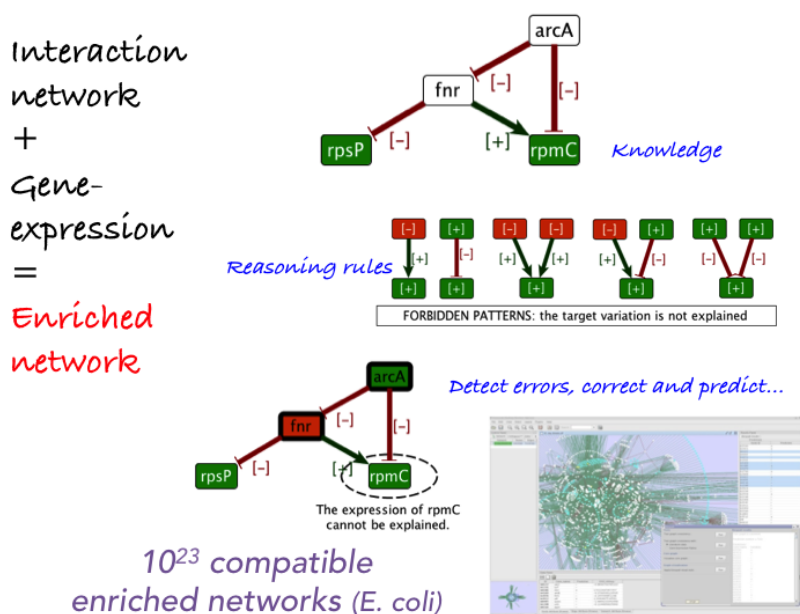


Figure 1.

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. The ASP-based logical approach is flexible enough to model in a single framework network characteristics (products, interactions, partial information on signs of regulations and observations) and static rules about the effects of the dynamics of the system. Extensions of this framework include the exhaustive search for system repair or more constrained dynamical rules. [5], [4]

**Step 1.** Regulation knowledge is represented as a signed oriented graph. Edge colors stand for regulatory effects (red/green  $\rightarrow$  inhibition or activation). Vertex colors stand for gene expression data (red/green  $\rightarrow$  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 inconsistent with the regulation in the graph).

### 3.2. Probabilistic and symbolic dynamics

We work on new techniques to emphasize biological strategies that must occur to reproduce quantitative measurements in order to predict the quantitative response of a system at a larger-scale. Our framework mixes mechanistic and probabilistic modeling [1]. 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 [52]. 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.

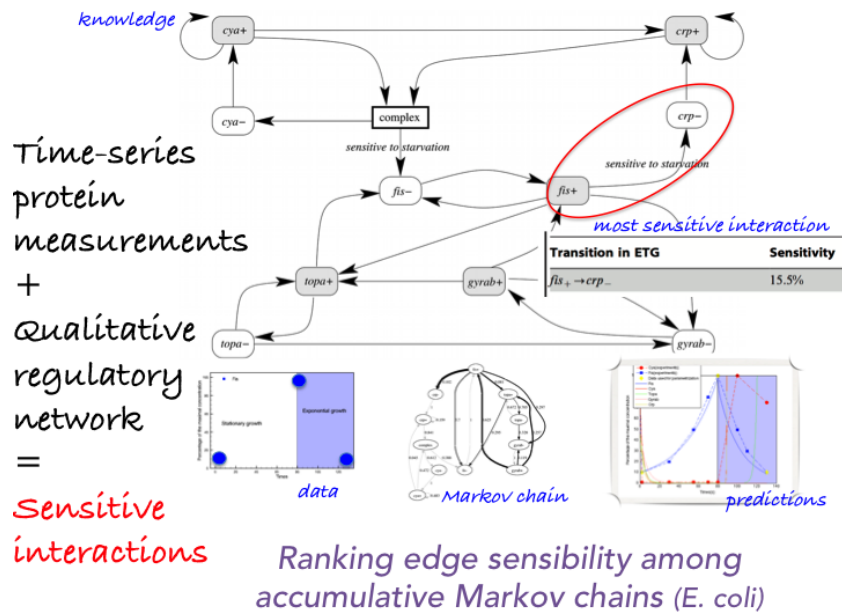


Figure 2.

Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems and identification of key interactions [1].

**Input data** are provided by 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. Estimation by local searches in the space 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 fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in *E. Coli* response to nutritional stress.

### 3.3. Grammatical inference and highly expressive structures

Our main field of expertise in machine learning concerns grammatical models with a strong expertise in finite state automata learning. 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 [3], leading to a tool named Protomata-learner. As an example, this tools allows us to properly model the function of the protein family TNF, which is impossible with other existing probabilistic-based approach (see Fig. 3 ). It was also applied to model families of proteins in cyanobacteria [2]. Our future goal

is to further demonstrate the relevance of formal language modelling by addressing the question of enzyme prediction, from their genomic or protein sequences, aiming at better sensitivity and specificity. As enzyme-substrate interactions are very specific central relations for integrated genome/metabolome studies and are characterized by faint signatures, we shall rely on models for active sites involved in cellular regulation or catalysis mechanisms. This requires to build models gathering both structural and sequence information in order to describe (potentially nested or crossing) long-term dependencies such as contacts of amino-acids that are far in the sequence but close in the 3D protein folding. We wish to extend our expertise towards inferring Context-Free Grammars including the topological information coming from the structural characterization of active sites.

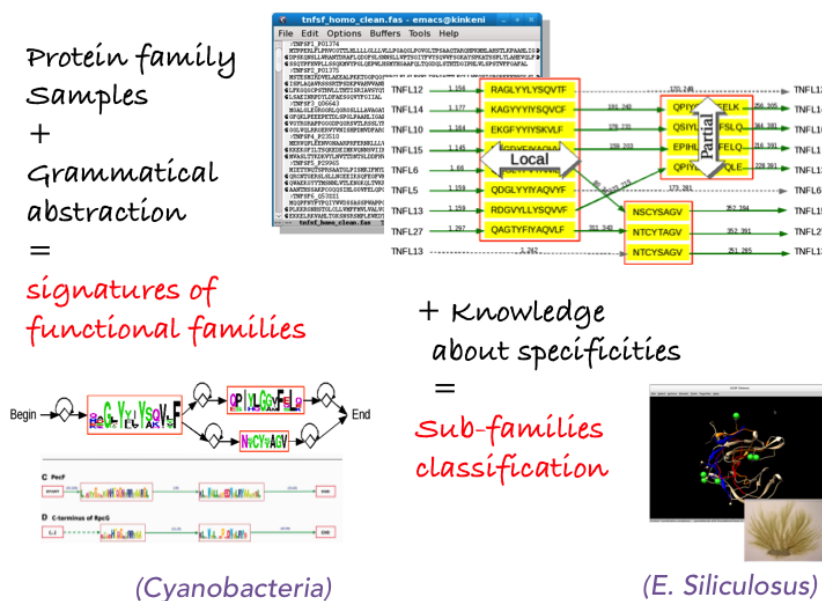


Figure 3. **Protomata Learner workflow.** Starting from a set of protein sequences, a partial local alignment is computed and an automaton is inferred, which can be considered as a signature of the family of proteins. This allows searching for new members of the family [2]. Adding further information 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 [54]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [59], [58]. 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 tool, STAN (suffix-tree analyser) which makes it possible to search for a subset of SVG patterns in full chromosome sequences [7]. This tool was used for the recognition of transposable elements in *Arabidopsis thaliana* [60] or for the design of a CRISPR database [9]. See Figure 4 for illustration. Our goal is 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, a language and framework based on a systematic introduction of constraints on string variables.

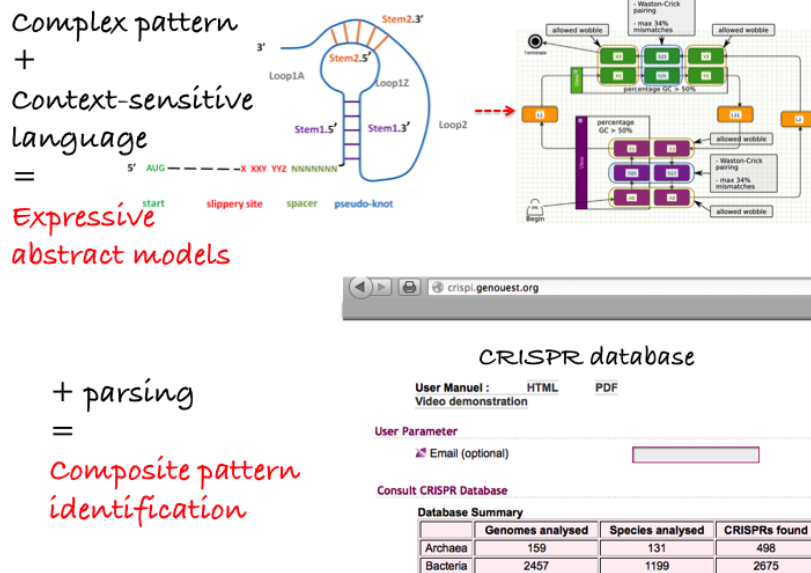


Figure 4. Graphical modeling of a pseudo-knot (RNA structure) based on the expressivity of String Variable Grammars used in the Logol framework. Combined with parsers, this leads to composite pattern identification such as CRISPR [56].

## GENSCALE Project-Team

### 3. Research Program

#### 3.1. Introduction

To tackle challenges brought by the processing of huge amount of genomic data, the main strategy of GenScale is to merge the following computer science expertise:

- Data structure;
- Combinatorial optimization;
- Parallelism.

#### 3.2. Data structure

To face the genomic data tsunami, the design of efficient algorithms involves the optimization of memory footprints. A key point is the design of innovative data structures to represent large genomic datasets into computer memories. Today's limitations come from their size, their construction time, or their centralized (sequential) access. Random accesses to large data structures poorly exploit the sophisticated processor cache memory system. New data structures including compression techniques, probabilistic filters, approximate string matching, or techniques to improve spatial/temporal memory access are developed [3].

#### 3.3. Combinatorial optimization

For wide genome analysis, Next Generation Sequencing (NGS) data processing or protein structure applications, the main issue concerns the exploration of sets of data by time-consuming algorithms, with the aim of identifying solutions that are optimal in a predefined sense. In this context, speeding up such algorithms requires acting on many directions: (1) optimizing the search with efficient heuristics and advanced combinatorial optimization techniques [2], [5] or (2) targeting biological sub-problems to reduce the search space [7], [9]. Designing algorithms with adapted heuristics, and able to scale from protein (a few hundreds of amino acids) to full genome (millions to billions of nucleotides) is one of the competitive challenges addressed in the GenScale project.

#### 3.4. Parallelism

The traditional parallelization approach, which consists in moving from a sequential to a parallel code, must be transformed into a direct design and implementation of high performance parallel software. All levels of parallelism (vector instructions, multi-cores, many-cores, clusters, grid, clouds) need to be exploited in order to extract the maximum computing power from current hardware resources [6], [8], [1]. An important specificity of GenScale is to systematically adopt a design approach where all levels of parallelism are potentially considered.

## 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 (hyper)rectangular regions in the state space and transitions between states arise from solutions entering one region from another.

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

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

**Participants:** Eugenio Cinquemani [Correspondent], Johannes Geiselmann, Hidde de Jong, Julien Demol, Stéphan Lacour, Michel Page, Corinne Pinel, Delphine Ropers, Alberto Soria-Lopéz, Diana Stefan, 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.

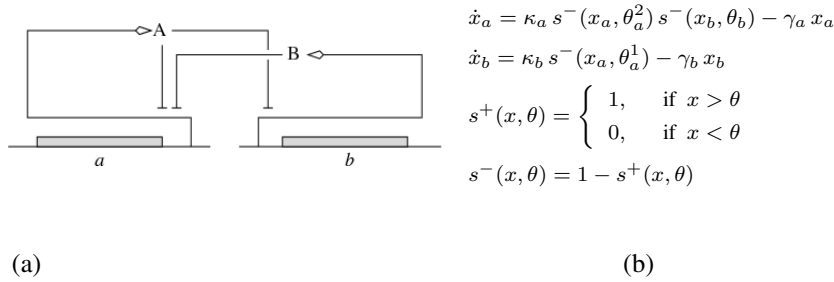


Figure 2. (Left) Example of a gene regulatory network of two genes ( $a$  and  $b$ ), each coding for a regulatory protein ( $A$  and  $B$ ). Protein  $B$  inhibits the expression of gene  $a$ , while protein  $A$  inhibits the expression of gene  $b$  and its own gene. (Right) PLDE model corresponding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

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 (Section 4.2). 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], Diana Stefan, 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

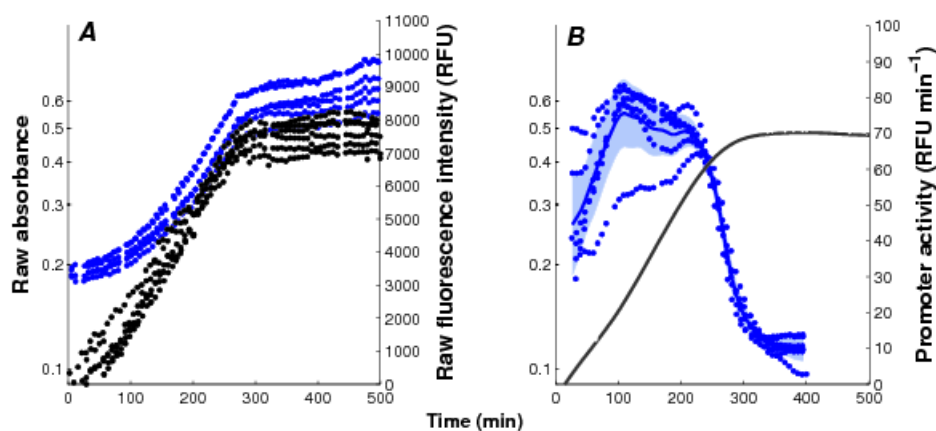


Figure 3. Monitoring of bacterial gene expression in vivo using fluorescent reporter genes (Stefan et al., *PLoS Computational Biology*, 11(1):e1004028, 2015). The plots show the primary data obtained in a kinetic experiment with *E. coli* cells, focusing on the expression of the motility gene *tar*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

control metabolic fluxes, while the accumulation or depletion of metabolites may affect the activity of transcription factors and thus the expression of enzyme-encoding genes.

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

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

### 3.4. Natural and engineered control of regulatory networks



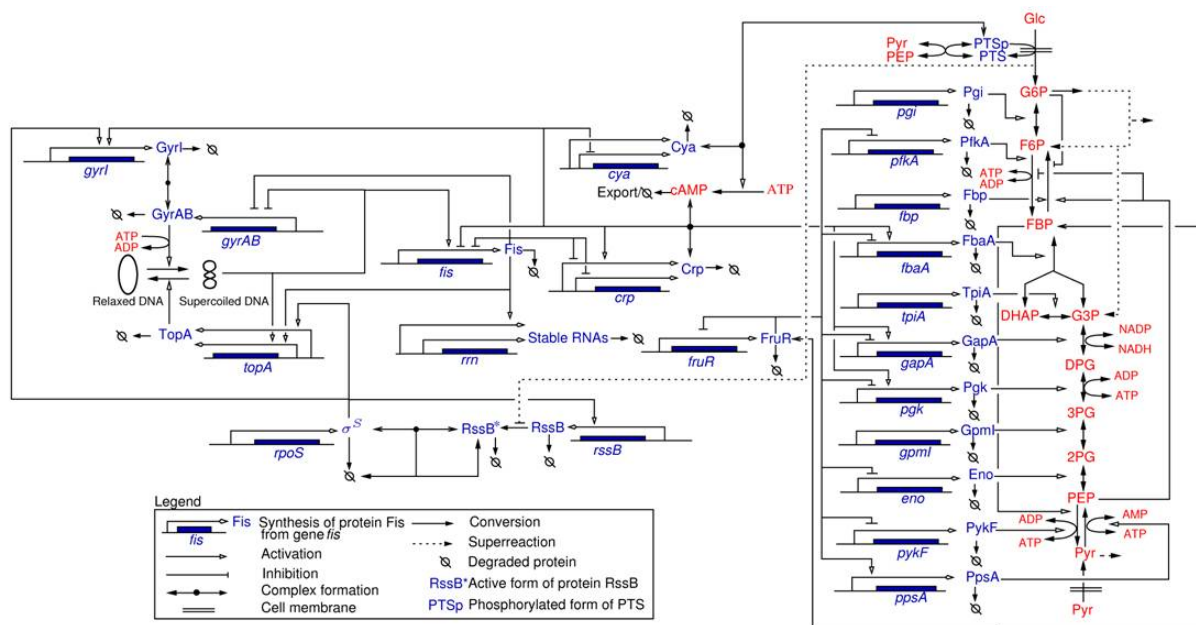


Figure 4. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in *E. coli* (Baldazzi et al., *PLoS Computational Biology*, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description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

**Participants:** Cindy Gomez Balderas-Barillot, Eugenio Cinquemani, Johannes Geiselman [Correspondent], Edith Grac, Nils Giordano, Hidde de Jong, Stéphan Lacour, Delphine Ropers, Alberto Soria-Lopéz.

In the previously-described objectives, we have focused on identifying complex regulatory networks and gaining a better understanding of how the network dynamics underlies the observable behavior of the cell. Based on the insights thus obtained, a complementary perspective consists in changing 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 growth-rate controllers of bacterial cells. Since the growth rate is the most important physiological parameter in microorganisms, a better understanding of the molecular basis of growth-rate control and the engineering of open-loop and closed-loop growth-rate controllers is of major interest for both fundamental research and biotechnological applications. Second, we are working on the development of methods with a solid foundation in control theory 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 CONTRAINTES, in collaboration with a biophysics group at Université Paris Descartes.

## LIFEWARE 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 <sup>0</sup>.

### 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 <sup>0</sup>, the inference of reaction systems from ODEs [8], the computation of structural invariants by constraint programming techniques, and the analysis of model reductions by subgraph epimorphisms [9], 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) <sup>0</sup>. The logical paradigm for systems biology that we have subsequently developed for quantitative models can be summarized by the following identifications :

$$\begin{aligned} \text{biological model} &= \text{transition system,} \\ \text{biological property} &= \text{temporal logic formula,} \\ \text{model validation} &= \text{model-checking,} \\ \text{model inference} &= \text{constraint solving.} \end{aligned}$$

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 [20], [19] 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**) <sup>0</sup> has shown to provide best results in our context, for

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

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

<sup>0</sup>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.

<sup>0</sup>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<sup>0</sup>

---

<sup>0</sup>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  $\beta$ -arrestin signaling. *Molecular Systems Biology*, 8(590), 2012.

## MAGNOME Project-Team

### 3. Research Program

#### 3.1. Overview

Fundamental questions in the life sciences can now be addressed at an unprecedented scale through the combination of high-throughput experimental techniques and advanced computational methods from the computer sciences. The new field of *computational biology* or *bioinformatics* has grown around intense collaboration between biologists and computer scientists working towards understanding living organisms as *systems*. One of the key challenges in this study of systems biology is understanding how the static information recorded in the genome is interpreted to become dynamic systems of cooperating and competing biomolecules.

MAGNOME addresses this challenge through the development of informatic techniques for understanding the structure and history of eukaryote genomes: algorithms for genome analysis, data models for knowledge representation, stochastic hierarchical models for behavior of complex systems, and data mining and classification. Our work is in methods and algorithms for:

- **Genome annotation** for complete genomes, performing *syntactic* analyses to identify genes, and *semantic* analyses to map biological meaning to groups of genes [22], [5], [9], [10], [32], [33].
- **Integration of heterogeneous data**, to build complete knowledge bases for storing and mining information from various sources, and for unambiguously exchanging this information between knowledge bases [1], [3], [25], [27], [21].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [10], [7], [28], [34].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [8], [6], [9].
- **Hierarchical and comparative modeling**, to build mathematical models of the behavior of complex biological systems, in particular through combination, reutilization, and specialization of existing continuous and discrete models [24], [20][12].

The hundred- to thousand-fold decrease in sequencing costs seen in the past few years presents significant challenges for data management and large-scale data mining. MAGNOME's methods specifically address "scaling out," where resources are added by installing additional computation nodes, rather than by adding more resources to existing hardware. Scaling out adds capacity and redundancy to the resource, and thus fault tolerance, by enforcing data redundancy between nodes, and by reassigning computations to existing nodes as needed.

#### 3.2. Comparative genomics

The central dogma of evolutionary biology postulates that contemporary genomes evolved from a common ancestral genome, but the large scale study of their evolutionary relationships is frustrated by the unavailability of these ancestral organisms that have long disappeared. However, this common inheritance allows us to discover these relationships through *comparison*, to identify those traits that are common and those that are novel inventions since the divergence of different lineages.

We develop efficient methodologies and software for associating biological information with complete genome sequences, in the particular case where several phylogenetically-related eukaryote genomes are studied simultaneously.

The methods designed by MAGNOME for comparative genome annotation, structured genome comparison, and construction of integrated models are applied on a large scale to: eukaryotes from the hemiascomycete class of yeasts [32], [33], [5], [9], [2], [10] and to lactic bacteria used in winemaking [31], [26]

### 3.3. Comparative modeling

A general goal of systems biology is to acquire a detailed quantitative understanding of the dynamics of living systems. Different formalisms and simulation techniques are currently used to construct numerical representations of biological systems, and a recurring challenge is that hand-tuned, accurate models tend to be so focused in scope that it is difficult to repurpose them. We claim that, instead of modeling individual processes *de novo*, a sustainable effort in building efficient behavioral models must proceed incrementally. *Hierarchical modeling* 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 combined uses theoretical results from formal methods and practical considerations from modeling applications to define a framework in which discrete and continuous models can communicate with a clear semantics. Hierarchical models can be assembled from existing models, and translated into their execution semantics and then simulated at multiple resolutions through multi-scale stochastic simulation. These models are compiled into a discrete event formalism capable of capturing discrete, continuous, stochastic, non deterministic and timed behaviors in an integrated and non-ambiguous way. Our long-term goal to develop a methodology in which we can **assemble a model** for a species of interest using a library of reusable models and a organism-level “schematic” determined by comparative genomics.

Comparative modeling is also a matter of reconciling experimental data with models [4] [20] and inferring new models through a combination of comparative genomics and successive refinement [29], [30].

## 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 experiments generally yield huge data sets resulting from high throughput screenings. The research of Morpheme will be developed along the following axes:

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

## 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 [6], [2], [7], [4]. This idea has been successfully applied to wide-field, spinning-disk confocal microscopy [1], TIRF [29], 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 5.1) 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 [43], PALM [31]) SIM has the strong advantage of versatility when considering the photo-physical properties of the fluorescent probes [40]. 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 [52]. Data association even combined with sophisticated particle filtering techniques [55] or matching techniques [53] 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 [36], [41]. 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 [55], [39]. Among the methods proposed to detect particles in microscopy images [57], [54], 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/disappearances 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 [34]). 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 [44], [59] 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.uhc.edu/>). Previous simulation-based methods have been investigated to explain the spatial organization of microtubules [47] 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) [58] 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. A difficulty is related to the scale change and statistical aggregation problems (in time and space) have to be handled.

## VIRTUAL PLANTS Project-Team

### 3. Research Program

#### 3.1. Analysis of structures resulting from meristem activity

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

- *Combinatorial approaches*: plant structures exhibit complex branching organizations of their organs like internodes, leaves, shoots, axes, branches, etc. These structures can be analyzed with combinatorial methods in order to compare them or to reveal particular types of organization. We investigate a family of techniques to quantify distances between branching systems based on non-linear structural alignment (similar to edit-operation methods used for sequence comparison). Based on these techniques, we study the notion of (topology-based) self-similarity of branching structures in order to define a notion of degree of redundancy for any tree structure and to quantify in this way botanical notions, such as the physiological states of a meristem, fundamental to the description of plant morphogenesis.
- *Statistical modeling*: We investigate different categories of statistical models corresponding to different types of structures.
  - Longitudinal data corresponding to plant growth follow up: the statistical models of interest are equilibrium renewal processes 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.

## 3.3. OpenAlea: An open-software platform for plant modeling

*OpenAlea* is open-software platform for interdisciplinary research in plant modeling and simulation. This scientific workflow platform is used for the integration and comparison of different models and tools provided by the research community. It is based on the Python (<http://www.python.org>) language that aims at being both a *glue* language for the different modules and an efficient modeling language for developing new models and tools. *OpenAlea* currently includes modules for plant simulation, analysis and modeling at different scales (*V-Plants* modules), for modeling ecophysiological processes (*Alinea* modules) such as radiative transfer, transpiration and photosynthesis (*RATP*, *Caribu*, *Adel*, *TopVine*, *Ecomeristem*) and for 3D visualization of plant architecture at different scales (*PlantGL*).

*OpenAlea* is the result of a collaborative effort associating 20 french research teams in plant modeling from Inria, CIRAD, INRA and ENS Lyon. The Virtual Plants team coordinates both development and modeling consortiums, and is more particularly in charge of the development of the kernel and of some of the main data structures such as multi-scale tree graphs and statistical sequences.

*OpenAlea* is a fundamental tool to share models and methods in interdisciplinary research (comprising botany, ecophysiology, forestry, agronomy, applied mathematics and computer science approaches). Embedded in Python and its scientific libraries, the platform may be used as a flexible and useful toolbox by biologists and modelers for various purposes (research, teaching, rapid model prototyping, communication, etc.).

## **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 [56]. 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 [48], [46]. Regarding target applications, a good review of the state of the art can be found in the book *Computer Integrated Surgery* [44], in N. Ayache's article [51] and in the more recent syntheses [52] [56]. The scientific journals *Medical Image Analysis* [39], *Transactions on Medical Imaging* [45], and *Computer Assisted Surgery* [47] are also good reference material. One can have a good vision of the state of the art with the proceedings of the most recent conferences MICCAI'2010 (Medical Image Computing and Computer Assisted Intervention) [42], [43] or ISBI'2010 (Int. Symp. on Biomedical Imaging) [41].

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 [57], [67]. It is also possible to obtain from a Magnetic Resonance Image (MRI) of the head a reasonable segmentation into skull tissues, white matter, grey matter, and cerebro-spinal fluid [70], or to measure some functional properties of the heart from dynamic sequences of Magnetic Resonance [50], Ultrasound or Nuclear Medicine images [58].

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 physical modeling of the image acquisition process, only a few actually model the physical or even 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 comprehension 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 multi-sequence (or multi-parametric)<sup>0</sup> and multi-modal images<sup>0</sup> for each single patient.

---

<sup>0</sup>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 Magnetic Resonance Imaging (MRI), patients followed for multiple sclerosis may undergo every six months a 3-D 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 MR images 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.

<sup>0</sup>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, MR providing contrast 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:

1. Multi-dimensional, multi-sequence and multi-modal image segmentation,
2. Image Registration/Fusion,

### 3.3. Computational Anatomy

The objective of the Computational Anatomy (CA) is the modeling and analysis of biological variability of 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<sup>0</sup>.

Studying the variability of biological shapes is an old problem (cf. the remarkable book "On Shape and Growth" by D'Arcy Thompson [69]). Significant efforts have been made since that time to develop a theory for statistical shape analysis (one can refer to [55] for a good synthesis, and to the special issue of Neuroimage [68] for recent developments). Despite all these efforts, there are a number of challenging mathematical issues which remain largely unsolved in general. A particular issue is the computation of statistics on manifolds which can be of infinite dimension (e.g the group of diffeomorphisms).

There is a classical stratification of the problems into the following 3 levels [64]: 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 to the following problems:

1. Statistics on anatomical manifolds,
2. Propagation of variability from anatomical manifolds,
3. Linking anatomical variability to image analysis algorithms,
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 be used for instance 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.

Quite advanced models have already been proposed to study at the molecular, cellular and organic level a number of physiological systems (see for instance [65], [62], [53], [66], [59]). 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.

---

<sup>0</sup>The NIH has launched 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 objective will be to establish new surrogate end-points from the automated analysis of temporal sequences. This is a challenging objective for researchers in Computational Anatomy. The data will be made available to qualified research groups involved or not in the study.

There is a hierarchy of modeling levels for CP models of the human body [54]:

- the first level is mainly geometrical, and addresses the construction of a digital description of the anatomy [49], essentially acquired from medical imagery;
- the second level is physical, involving mainly the biomechanical modeling of various tissues, organs, vessels, muscles or bone structures [60];
- the third level is physiological, involving a modeling of the functions of the major organic systems [61] (e.g. cardiovascular, respiratory, digestive, central or peripheral nervous, muscular, reproductive, hormonal, etc.) or some pathological metabolism (e.g. evolution of cancerous or inflammatory lesions, formation of vessel stenoses, etc.);
- a fourth level would be cognitive, modeling the higher functions of the human brain [40].

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 [63]). 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 first level of modeling) to a much more ambitious *Physiological Human project* (see [61], [62]). We will not address all the issues raised by this ambitious project, but instead focus on topics detailed below. Among them, our objective is to identify some common methods for the resolution of the large inverse problems raised by the coupling of physiological models to medical images for the construction of patient-specific models (e.g. specific variational or sequential methods (EKF), dedicated particle filters, etc.). We also plan to develop specific expertise on 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, etc.). Application domains include

1. Surgery Simulation,
2. Cardiac Imaging,
3. Brain tumors, neo-angiogenesis, wound healing processes, ovocyte regulation, ...

### 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 demonstration of feasibility on a limited number of representative examples (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 is also often the occasion to get access to larger databases of images and signals which in turn help stimulate of 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 [64], Merboldt et al [68] and Taylor et al [80]. 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. The great success of dMRI comes from its ability to accurately describe the geometry of the underlying microstructure and probe the structure of the biological tissue at scales much smaller than the imaging resolution.

The diffusion of water molecules is Brownian in an isotropic medium and under normal unhindered conditions, but in fibrous structure such as white matter, the diffusion is very often directionally biased or anisotropic and water molecules tend to diffuse along fibers. For example, a molecule inside the axon of a neuron has a low probability to cross a myelin membrane. Therefore the molecule will move principally along the axis of the neural fiber. Conversely if we know that molecules locally diffuse principally in one direction, we can make the assumption that this corresponds to a set of fibers.

##### 3.1.1. Diffusion Tensor Imaging

Shortly after the first acquisitions of diffusion-weighted images (DWI) were made in vivo [70], [71], Basser et al [45], [44] proposed the rigorous formalism of the second order Diffusion Tensor Imaging model (DTI). DTI describes the three-dimensional (3D) nature of anisotropy in tissues by assuming that the average diffusion of water molecules follows a Gaussian distribution. It encapsulates the diffusion properties of water molecules in biological tissues (inside a typical  $1\text{-}3\text{ mm}^3$  sized voxel) as an effective self-diffusion tensor given by a  $3 \times 3$  symmetric positive definite tensor  $\mathbf{D}$  [45], [44]. Diffusion tensor imaging (DTI) thus produces a three-dimensional image containing, at each voxel, the estimated tensor  $\mathbf{D}$ . This requires the acquisition of at least six Diffusion Weighted Images (DWI)  $S_k$  in several non-coplanar encoding directions as well as an unweighted image  $S_0$ . Because of the signal attenuation, the image noise will affect the measurements and it is therefore important to take into account the nature and the strength of this noise in all the pre-processing steps. From the diffusion tensor  $\mathbf{D}$ , a neural fiber direction can be inferred from the tensor's main eigenvector while various diffusion anisotropy measures, such as the Fractional Anisotropy (FA), can be computed using the associated eigenvalues to quantify anisotropy, thus describing the inequality of diffusion values among particular directions.

DTI has now proved to be extremely useful to study the normal and pathological human brain [65], [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 driven our interest to develop new processing tools for Diffusion MRI. Because of the complexity of the data, this imaging modality raises a large amount of mathematical and computational challenges. We have therefore started to develop 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 [67], [8] and [66]).

### 3.1.2. High Angular Resolution Diffusion Imaging

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. This is an important limitation, since resolution of DTI acquisition is between  $1\text{mm}^3$  and  $3\text{mm}^3$  while the physical diameter of fibers can be between  $1\mu\text{m}$  and  $30\mu\text{m}$  [76], [46]. Research groups currently agree that there is complex fiber architecture in most fiber regions of the brain [75]. In fact, it is currently thought that between one third to two thirds of imaging voxels in the human brain white matter contain multiple fiber bundle crossings [47]. This has led to the development of various High Angular Resolution Diffusion Imaging (HARDI) techniques [82] such as Q-Ball Imaging (QBI) or Diffusion Spectrum Imaging (DSI) [83], [84], [86] to explore more precisely the microstructure of biological tissues.

HARDI samples q-space along as many directions as possible in order to reconstruct estimates of the true diffusion probability density function (PDF) – also referred as the Ensemble Average Propagator (EAP) – of water molecules. This true diffusion PDF is model-free and can recover the diffusion of water molecules in any underlying fiber population. HARDI depends on the number of measurements  $N$  and the gradient strength ( $b$ -value), which will directly affect acquisition time and signal to noise ratio in the signal.

Typically, there are two strategies used in HARDI: 1) sampling of the whole q-space 3D Cartesian grid and estimation of the EAP by inverse Fourier transformation or 2) single shell spherical sampling and estimation of fiber distributions from the diffusion/fiber ODF (QBI), Persistent Angular Structure [63] or Diffusion Orientation Transform [88]. In the first case, a large number of q-space points are taken over the discrete grid ( $N > 200$ ) and the inverse Fourier transform of the measured Diffusion Weighted Imaging (DWI) signal is taken to obtain an estimate of the diffusion PDF. This is Diffusion Spectrum Imaging (DSI) [86], [83], [84]. The method requires very strong imaging gradients ( $500 \leq b \leq 20000 \text{ s/mm}^2$ ) and a long time for acquisition (15-60 minutes) depending on the number of sampling directions. To infer fiber directions of the diffusion PDF at every voxel, people take an isosurface of the diffusion PDF for a certain radius. Alternatively, they can use the second strategy known as Q-Ball imaging (QBI) i.e just a single shell HARDI acquisition to compute the diffusion orientation distribution function (ODF). With QBI, model-free mathematical approaches can be developed to reconstruct the angular profile of the diffusion displacement probability density function (PDF) of water molecules such as the ODF function which is fundamental in tractography due to the fact that it contains the full angular information of the diffusion PDF and has its maxima aligned with the underlying fiber directions at every voxel.

QBI and the diffusion ODF 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 available. 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 [52], [4] and [53],[5]).

### 3.1.3. High Order Tensors

Other 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 [87], [89] or 4th order Tensor Model [43]. For more details, we refer the reader to our articles in [56], [79] where we review HOT models and to our articles in [7], 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) [59].

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. 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], [79].

### 3.1.4. Improving dMRI Acquisitions and Modeling

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 [69].

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 [69], [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 [69], [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<sup>0</sup>, have been selected for use in the HCP (Human Connectome Project)<sup>0</sup>.

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]

## 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

---

<sup>0</sup><http://www.emmanuelcaruyer.com/>

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

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 ODYSSEE/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 [6], [9] and means to calibrate them [85] 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 Fraise, Mitsuhiro Hayashibe, David Andreu.

We aim at developing realistic solutions for real clinical problems expressed by patients and medical staff. Different approaches and specifications are developed 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 developed in the third axis (Neuroprostheses). The robotics framework involved in DEMAR work is close to the tools used and developed 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 too 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:

$$Y^* = \operatorname{argmax}_Y s(X, Y) \quad (1)$$

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 .

Applying this framework requires (a) identifying which symbols and relations to use for image and object representation (b) learning a scoring function  $s$  from training data and (c) optimizing over  $Y$  in Eq. 1 . One of the main themes of our work is the development of methods that jointly address (a,b,c) in a shape-grouping framework in order to reliably extract, describe, model and detect shape information from natural and medical images. A principal motivation for using a shape-based framework is the understanding that shape- and more generally, grouping- based representations can go all the way from image features to objects. Regarding aspect (a), image representation, we cater for the extraction of image features that respect the shape properties of image structures. Such features are typically constructed to be purely geometric (e.g. boundaries, symmetry axes, image segments), or appearance-based, such as image descriptors. The use of machine learning has been shown to facilitate the robust and efficient extraction of such features, while the grouping of local evidence is known to be necessary to disambiguate the potentially noisy local measurements. In our research we have worked on improving feature extraction, proposing novel blends of invariant geometric- and appearance- based features, as well as grouping algorithms that allow for the efficient construction of optimal assemblies of local features.

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

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

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

### 3.2. Machine Learning & Structured Prediction

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

$$\mathcal{R}(f) = \int \ell(f, x, y) dP(x, y), \quad (2)$$

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,  $\ell$ , is informed by the costs associated with making a specific misprediction. As a concrete example, if the true spatial extent of a tumor is encoded in  $y$ ,  $f(x)$  may make mistakes in classifying healthy tissue as a tumor, and mistakes in classifying diseased tissue as healthy. The loss function should encode the potential physiological damage resulting from erroneously targeting healthy tissue for irradiation, as well as the risk from missing a portion of the tumor.

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

$$\mathcal{R}(f) \approx \widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, x_i, y_i), \quad (3)$$

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

$$\min_{f \in \mathcal{F}} \lambda \Omega(\|f\|) + \widehat{\mathcal{R}}(f), \quad (4)$$

where  $\Omega$  is a monotonically increasing function that penalizes complex functions.

Equation (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 (3). Furthermore, we typically gain performance by appropriately modeling the inter-relationships between voxel predictions, e.g. by incorporating pairwise and higher order potentials that encode prior knowledge about the problem domain. It is in this context that we develop statistical methods appropriate to structured prediction in the medical imaging setting.

### 3.3. Self-Paced Learning with Missing Information

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

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

$$\min_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w}) + \sum_{i=1}^n \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \quad (5)$$

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

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

$$\min_{\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \{0,1\}^n} R(\mathbf{w}) + \sum_{i=1}^n v_i \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})) - \mu_t \sum_{i=1}^n v_i. \quad (6)$$

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

$$\min_{\mathbf{w} \in \mathcal{W}, \theta \in \Theta} R(\mathbf{w}) + \sum_{i=1}^n \sum_{h_i \in \mathcal{H}} \Pr(h_i | x_i, y_i; \theta) \Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \quad (7)$$

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

### 3.4. Discrete Biomedical Image Perception

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

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

$$\arg \min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}). \quad (8)$$

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

$$\arg \min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}) + \sum_{C_i \in \mathcal{E}} f_{p_1 \dots p_n}(x_{p_{i^1}}, \dots, x_{p_{i^{|C_i|}}}). \quad (9)$$

where  $f_{p_1 \dots p_n}$  is the price to pay for associating the labels  $(x_{p_{i^1}}, \dots, x_{p_{i^{|C_i|}}})$  to the nodes  $(p_1 \dots p_{i^{|C_i|}})$ . Parameter inference, addressed by minimizing the problem above, is the most critical aspect in computational medicine and efficient optimization algorithms are to be evaluated both in terms of computational complexity as well as of inference performance. State of the art methods include deterministic and non-deterministic annealing, genetic algorithms, max-flow/min-cut techniques and relaxation. These methods offer certain strengths while exhibiting certain limitations, mostly related to the amount of interactions which can be tolerated among neighborhood nodes. In the area of medical imaging where domain knowledge is quite strong, one would expect that such interactions should be enforced at the largest scale possible.

## 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 in space and functional mechanisms in time. From a spatial 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 temporal 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.

Three important characteristics are worth mentioning concerning these loops. Firstly, each of them sets a closed relation between the central nervous system and the rest of the world. This includes the external world (possibly including other intelligent agents), but also the internal world, with hormonal, physiological and bodily dimensions. Secondly, each of these loops can be described as a loop relating sensations to actions, in the wide sense of these terms: effectively, action can refer to acting in the real world, but also to modifying physiological parameters or controlling neuronal activation. These loops have different constants of time, from immediate reflexes and sensorimotor adjustments to long term selection of motivation for action, the latter depending on hormonal and social parameters. Thirdly, each of the loops performs a learning reinforced by a primary (physiologically significant) or pseudo reward (sub-goal to be learned). As an illustration, we can mention respondent conditioning detecting stimuli anticipatory of primary rewards, episodic learning detecting multimodal events, and also more local phenomena like self-organization of topological structures. The gradual establishment of these loops and their mutual interactions give an interpretation of the resulting cognitive architecture as a synergetic system of memories.

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 [42]. 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 [38], most of these frameworks (e.g. Soar, ACT-R), though sometimes justified on biological grounds, do not go up to a *connectionist* neuronal implementation. Furthermore, because of the complexity of the resulting frameworks, they are restricted to simple symbolic interfaces with the internal and external world and to (relatively) small-sized internal structures. Our main research objective is undoubtedly to build such a general purpose cognitive architecture (to model the brain *as a whole* in a systemic way), using a connectionist implementation and able to cope with a realistic environment.

## 3.2. Computational Neuroscience

From a general point of view, computational neuroscience can be defined as the development of methods from computer science and applied mathematics, to explore more technically and theoretically the relations between structures and functions in the brain [44], [32]. 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 mimic biology through its most general law of homeostasis and self-adaptability in an unknown and changing environment. This means that we propose to numerically experiment such models and thus provide effective methods to falsify them.

Here, computational neuroscience means mimicking original computations made by the neuronal substratum and mastering their corresponding properties: computations are distributed and adaptive; they are performed without an homonculus or any central clock. Numerical schemes developed for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [29], [37] and were the basis for several contributions in our group [43], [40], [45]. 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 [30]; the associated formalism [35] 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 [27], that demonstrated a richness of behavior [31] 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 [40], [41] and their use for observing the emergence of typical cortical properties [34]. 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.* § 5.1 ). 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 § 5.1 ).

## 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 [36] 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 [33], [26]. Nevertheless, we believe that, as for previous models, machine learning algorithms remain isolated tools, whereas our systemic approach can bring original views on these problems.

At the cognitive level, most of the problems we face do not rely on only one kind of learning and require instead skills that have to be learned in preliminary steps. That is the reason why cognitive architectures are often referred to as systems of memory, communicating and sharing information for problem solving. Instead of the classical view in machine learning of a flat architecture, a more complex network of modules must be considered here, as it is the case in the domain of deep learning. In addition, our systemic approach brings the question of incrementally building such a system, with a clear inspiration from developmental sciences. In this perspective, modules can generate internal signals corresponding to internal goals, predictions, error signals, able to supervise the learning of other modules (possibly endowed with a different learning rule), supposed to become autonomous after an instructing period. A typical example is that of episodic learning (in the hippocampus), storing declarative memory about a collection of past episodes and supervising the training of a procedural memory in the cortex.

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

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

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

### 3.4. Autonomous Robotics

Autonomous robots are not only convenient platforms to implement our algorithms; the choice of such platforms is also motivated by theories in cognitive science and neuroscience indicating that cognition emerges



from interactions of the body in direct loops with the world and develops interesting specificities accordingly. For example, internal representations can be minimized (opposite to building complex and hierarchical representations) and compensated by more simple strategies [28], more directly coupling perception and action and more efficient to react quickly in the changing environment (for example, instead of memorizing details of an object, just memorizing the eye movement to foveate it: the world itself is considered as an external memory). In this view for the *embodiment of cognition*, learning is intrinsically linked to sensorimotor loops and to a real body interacting with a real environment.

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 [53], [57], [45], models of epilepsy [81], effects of synaptic plasticity [77], [78], homeostasis and intrinsic plasticity [8].

[Selected publications on this topic.](#)

#### 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  $50\mu\text{m}$  to 1mm, containing of the order of one hundred to one hundred thousand neurons belonging to a few different species. The description of this collective dynamics requires models which are different from individual neurons models. In particular, when the number of neurons is large enough averaging effects appear, and the collective dynamics is well described by an effective mean-field, summarizing the effect of the interactions of a neuron with the other neurons, and depending on a few effective control parameters. This vision, inherited from statistical physics requires that the space scale be large enough to include a large number of microscopic components (here neurons) and small enough so that the region considered is homogeneous.

Our group is developing mathematical and numerical methods allowing on one hand to produce dynamic mean-field equations from the physiological characteristics of neural structure (neurons type, synapse type and anatomical connectivity between neurons populations), and on the other so simulate these equations. These methods use tools from advanced probability theory such as the theory of Large Deviations [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. [67]) 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 [82]. 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.**

### 3.4. 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 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], [54], [45]. 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 [55]. On the other hand, we are proposing new algorithms for data processing [25]. 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.**

### 3.5. 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 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 [77], [78], [56]. 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 [63]. 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 [64].

[Selected publications on this topic.](#)

### 3.6. 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](#); [Institut de la vision, Paris](#); [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 [76], [79], [52], [9]. This work is done in collaboration with [Institut des Neurosciences de la Timone, Marseille](#).

[Selected publications on this topic.](#)

### 3.7. 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 [80], coding/decoding approaches [71], [72] or classification [60], [61].

[Selected publications on this topic.](#)

## NEUROSYS 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.

#### 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 focusses 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 noninvasively. 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  $\beta$  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  $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_f}$  of  $\mathbf{Y}$  independently, yielding  $n_f$  problems to be solved in parallel:

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

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

$$\hat{\beta}_i = \operatorname{argmin}_{\beta_i} \|\mathbf{Y}_i - \mathbf{X}\beta_i\|^2 + \Psi(\beta_i), \quad (10)$$

where  $\Psi$  is an adequate penalization used to regularize the solution:

$$\Psi(\beta; \lambda_1, \lambda_2, \eta_1, \eta_2) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2 + \eta_1 \|\nabla\beta\|_1 + \eta_2 \|\nabla\beta\|_2 \quad (11)$$

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

- When  $\lambda_1 > 0$  only (LASSO), and to some extent, when  $\lambda_1, \lambda_2 > 0$  only (elastic net), the optimal solution  $\beta$  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  $i \in \{1, \dots, n_f\}$ . This can be computed through cross-validation, by *learning*  $\hat{\beta}_i$  on some part of the dataset, and then estimating  $(Y_i - X\hat{\beta}_i)$  using the remainder of the dataset.

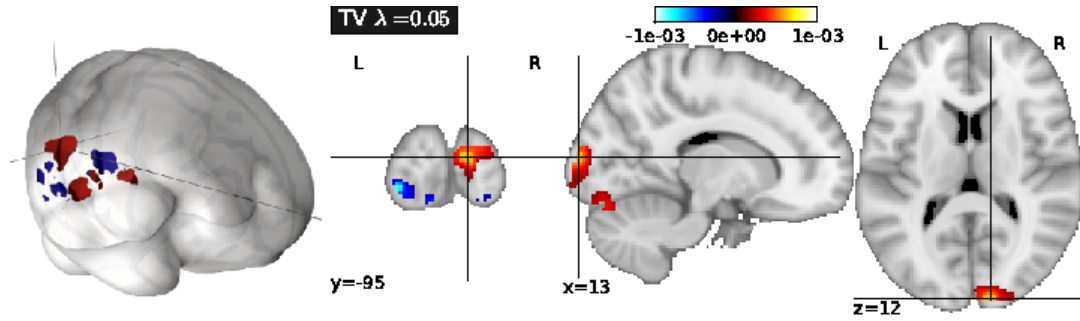


Figure 1. Example of the regularization of a brain map with total variation in an inverse problem. The problem here consists in predicting the spatial scale of an object presented as a stimulus, given functional neuroimaging data acquired during the observation of an image. Learning and test are performed 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  $\beta$  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  $\beta_i$  and thus, improve the prediction of  $\mathbf{Y}_i$ . Yet this does not impose constraints on the non-zero parameters of the parameters  $\beta_i$ .

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad (12)$$

then

$$\hat{\beta} = \operatorname{argmin}_{\beta=(\beta_i), i=1..n_f} \sum_{i=1}^{n_f} \|\mathbf{Y}_i - \mathbf{X}\beta_i\|^2 + \lambda \sum_{j=1}^{n_{\text{voxels}}} \sqrt{\sum_{i=1}^{n_f} \beta_{i,j}^2} \quad (13)$$



## 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

$$\mathbf{X} = \mathbf{A}\mathbf{D} + \epsilon, \quad (14)$$

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:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{A}\mathbf{D}\|^2 + \Psi(\mathbf{D}) \text{ s.t. } \|\mathbf{A}_i\| = 1 \forall i \in \{1..n_f\}, \quad (15)$$

where  $(\mathbf{A}_i)$ ,  $i \in \{1..n_f\}$  represents the columns of  $\mathbf{A}$ .  $\Psi$  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).

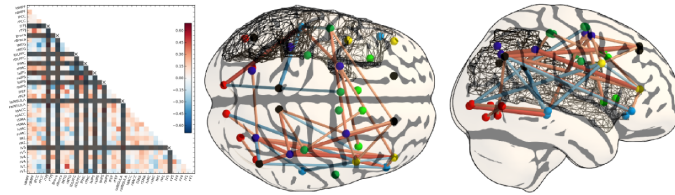
## 3.3. Covariance estimation

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.

## SHACRA Project-Team

### 3. Research Program

#### 3.1. Real-Time Biophysical Models

The principal objective of this scientific challenge is the modeling of the operative field, *i.e.* the anatomy and physiology of the patient that will be directly or indirectly targeted by a medical intervention. This requires to describe various phenomena such as soft-tissue deformation, fluid dynamics, electrical diffusion, or heat transfer. These models will help simulate the reaction of the patient's anatomy to the procedure, 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 fast, possibly real-time, computation.

##### 3.1.1. *Real-time biomechanical modeling of solid structures*

Soft tissue modeling holds a very important place in medical simulation. A large part of the realism of a simulation, in particular for surgery or laparoscopy simulation, relies upon the ability to describe soft tissue response during the simulated intervention. Several approaches have been proposed over the past ten years to model soft-tissue deformation in real-time (mainly for solid organs), usually based on elasticity theory and a finite element approach to solve the equations. We were among the first to propose an approach [3] using different computational strategies. Although significant improvements were obtained later on (for instance with the use of co-rotational methods to handle geometrical non-linearities) these works remain of limited clinical use as they essentially rely on linearized constitutive laws, and are rarely validated. An important part of our research remains dedicated to the development of new, more accurate models that are compatible with real-time computation. Such advanced models will not only permit to increase the realism of future training systems, but they will act as a bridge toward the development of patient-specific preoperative planning as well as augmented reality tools for the operating room.

##### 3.1.2. *Real-time biomechanical modeling of hollow structures*

A large number of anatomical structures in the human body are vascularized (brain, liver, heart, kidneys, etc.) and recent interventions (such as interventional radiology procedures) rely on the vascular network as a therapeutical pathway. It is therefore essential to model the shape and deformable behavior of blood vessels. This can be done at two levels, depending of the objective. The global deformation of a vascular network can be represented using the vascular skeleton as a deformable (tree) structure, while local deformations need to be described using models of deformable surfaces. Other structures such as aneurysms, the colon or stomach can also benefit from being modeled as deformable surface, and we can rely on shell or thin plate theory to reach this objective.

##### 3.1.3. *Coupled physical models*

Beyond biomechanical modeling of soft tissues, other physical phenomena have to be taken into account. In the context of percutaneous tumor ablation, both thermal and mechanical behaviors have to be modelled. Focusing especially on the simulation of cryoablation (freezing the pathological tissue), models for heat transfer have been implemented to simulate the evolution of temperature within living tissues. Pre-operative planning of the iceball can thus be envisaged. Moreover, this demonstrates the multi-physics aspect of SOFA.

##### 3.1.4. *Real-time electrophysiology*

Electrophysiology plays an important role in the physiology of the human body, for instance by inducing muscles motion, and obviously through the nervous system. Also, many clinical procedures rely on electrical stimulation, such as defibrillation, neuromuscular or deep brain stimulation for instance. Yet, the modeling and the simulation of this phenomenon is still in its early stages. Our primary objective is to focus on cardiac electrophysiology, which plays a critical role in the understanding of heart mechanisms, and also in the planning of certain cardiac procedures. We propose to develop models and computational strategies aimed at real-time simulation, and to also provide personalization tools (parameter estimation) allowing to run patient-specific simulations from clinical data.

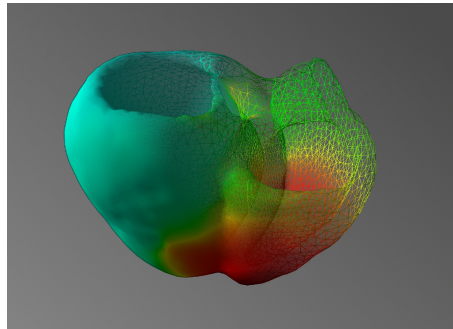


Figure 3. Patient-specific electrophysiology model of the human heart running in real-time

## 3.2. Numerical Methods for Complex Interactions

### 3.2.1. Dynamic topological changes

As mentioned previously, assisting the surgeon by providing either pre-operative planning or per-operative guidance assumes to increase the level of complexity and accuracy of our models, thus making the simulation more computationally-demanding. Innovative numerical methods must therefore be investigated. For instance, efforts are made to couple SOFA with CGoGN. CGoGN is a library based on combinatorial maps theory, specialized for representing and manipulating meshes. It is able to represent consistently objects of different dimensions composed of arbitrary cells (polygonal faces, polyhedral volumes). It provides an efficient way to explore the cells and their neighborhood; it allows to store data with the cells (both at execution time and compile time) and to efficiently modify the connectivity of the mesh even in highly dynamic cases. Adaptive meshing and efficient topological algorithm are thus available in SOFA.

### 3.2.2. Constraint models and boundary conditions

To simulate soft-tissue deformations accurately, the modeling technique must account for the intrinsic behavior of the modeled organ as well as for its biomechanical interactions with surrounding tissues or medical devices. While the biomechanical behavior of important organs (such as the brain or liver) has been studied extensively in the past, only few works exist dealing with the mechanical interactions between the anatomical structures. For tissue–tool interactions, most techniques rely on simple contact models, whereas advanced phenomena such as friction are rarely taken into account. While simplifications can produce plausible results in the case of interaction between the manipulator of a laparoscopic instrument and the surface of an organ, it is generally an insufficient approximation. As we move towards the simulations for planning or rehearsal, accurate modeling of contacts is playing an increasingly important role. For example, we have shown in [36] and [37] that complex interactions between a coil and an aneurysm, or alternatively between a flexible needle and a soft-tissue can be computed in real-time. In laparoscopic surgery, the main challenge is represented by modeling of interactions between anatomical structures rather than only between the instruments and the surface of the organ. Consequently, our objective was to model accurately the contacts with friction and other type on non-smooth interactions in a heterogeneous environment and to allow for stable haptic rendering. When different time integration strategies are used, another challenge is to compute the contact forces in such a way that integrity and stability of the overall simulation are maintained. Our objective was to propose a unified definition of such various boundary conditions and develop new numerical methods for simulations of heterogeneous objects.

## 3.3. Image-Driven Simulation: towards pre-operative planning and per-operative guidance

Image-guided therapy 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 typically related to image-guided therapy, such as multi-modality image registration, which serves to align pre-operative images onto the patient. As most procedures deal with soft-tissues, elastic registration techniques are necessary to perform this step. Novel registration techniques began to account for soft tissue deformation 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 time increases, thus lacking responsiveness. Second, as we have seen previously, many factors influence the deformation of soft-tissues, from patient-specific material properties to boundary conditions with surrounding anatomy. Another very similar, and related, problem is augmented reality, i.e. the real-time superposition of a virtual model onto the reality. In a clinical context, this can be very useful to help "see through" the anatomy. In this case, however, real-time registration of the virtual information onto the patient is mandatory. Our objective in this area is to combine our expertise in real-time soft-tissue modeling, complex interactions with image data to provide accurate and real-time registration, deformation, and tracking of virtual anatomical structures onto the patient.

The predictive capabilities of computer simulations may also be used to improve minimally invasive surgical procedures. While simulation results are sensitive to model parameters, initial and boundary conditions, we aim at combining computer-vision algorithms and simulation algorithms in order to produce dynamic data-driven simulation in clinical applications. The main idea is to use computer-vision algorithms from pre-operative diagnoses or per-operative video streams in order to extract meaningful data to feed the simulation engine and thus to increase the accuracy of the simulation. Clinical outcomes are expected in interventional radiology where the guidance is based on fluoroscopic imaging modality inducing high absorbed dose of X-rays for the patient and the clinical staff. In that context, using the prediction capabilities of the simulation may decrease the acquisition frequency of images, leading to a lower exposure of X-rays. Our objective in this area is to combine our expertise in patient-specific modeling and constraint models to achieve the dynamic coupling between images, pre-operative data and computer simulation.

## SISTM Team

### 3. Research Program

#### 3.1. Mechanistic modelling

When studying the dynamics of a given marker, say the HIV concentration in the blood (HIV viral load), one can for instance use descriptive models summarizing the dynamics over time in term of slopes of the trajectories [48]. 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 [41]. 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 [40].

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 [33]. 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 [32].

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 [45], [51], [36], [42], [37], [50], the difficulty associated with the complexity of ODE models and the sparsity of the data leading to identifiability issues need further research.

#### 3.2. High dimensional data

With the availability of omics data such as genomics (DNA), transcriptomics (RNA) or proteomics (proteins), but also other types of data, such as those arising from the combination of large observational databases (e.g. in pharmacoepidemiology or environmental epidemiology), high-dimensional data have become increasingly common. Use of molecular biological techniques such as Polymerase Chain Reaction (PCR) allows for amplification of DNA or RNA sequences. Nowadays, microarray and Next Generation Sequencing (NGS) techniques give the possibility to explore very large portions of the genome. Furthermore, other assays have also evolved, and traditional measures such as cytometry or imaging have become new sources of big data. Therefore, in the context of HIV research, the dimension of the datasets has much grown in term of number of variables per individual than in term of number of included patients although this latter is also growing thanks to the multi-cohort collaborations such as CASCADE or COHERE organized in the EuroCoord network<sup>0</sup>. As an example, in a recent phase 1/2 clinical trial evaluating the safety and the immunological response to a dendritic cell-based HIV vaccine, 19 infected patients were included. Bringing together data on cell count, cytokine production, gene expression and viral genome change led to a 20 Go database [19]. 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.

<sup>0</sup>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 [46]. 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 [49]. 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 [38]. 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.

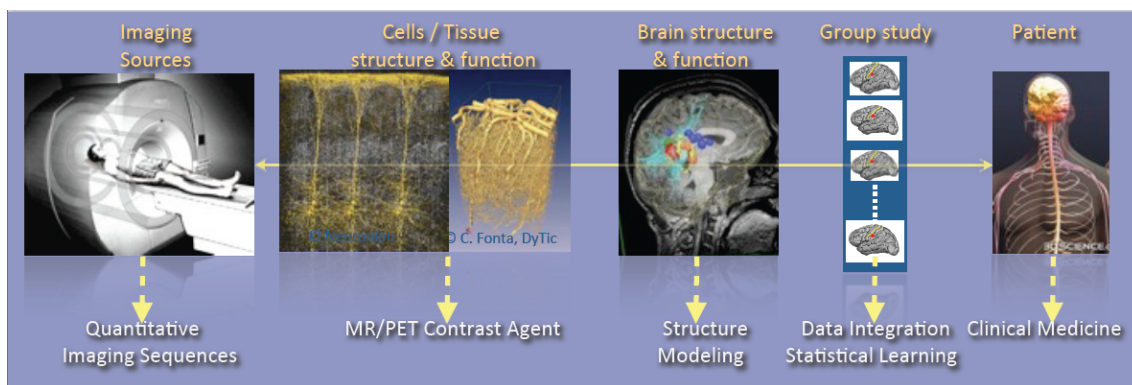


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales : from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions,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

## 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 [21], [24], [25]. 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 [22], [23] 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-physics systems are investigated. More details about the industrial impact of these studies are presented in the following section.

- To estimate the risk for infrastructures in coastal 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, Giorgio Giorgiani.

In order to fulfil the increasing demand, alternative energy sources have to be developed. Indeed, the current rate of fossil fuel usage and its serious adverse environmental impacts (pollution, greenhouse gas emissions, ...) lead to an energy crisis accompanied by potentially disastrous global climate changes.

Controlled fusion power is one of the most promising alternatives to the use of fossil resources, potentially with a unlimited source of fuel. France with the ITER (<http://www.iter.org/default.aspx>) and Laser Megajoule (<http://www-lmj.cea.fr/>) facilities is strongly involved in the development of these two parallel approaches to master fusion that are magnetic and inertial confinement. Although the principles of fusion reaction are well understood from nearly sixty years, (the design of tokamak dates back from studies done in the '50 by Igor Tamm and Andreï Sakharov in the former Soviet Union), the route to an industrial reactor is still long and the application of controlled fusion for energy production is beyond our present knowledge of related physical processes. In magnetic confinement, beside technological constraints involving for instance the design of plasma-facing component, one of the main difficulties in the building of a controlled fusion reactor is the poor confinement time reached so far. This confinement time is actually governed by turbulent transport that therefore determines the performance of fusion plasmas. The prediction of the level of turbulent transport in large machines such as ITER is therefore of paramount importance for the success of the researches on controlled magnetic fusion.

The other route for fusion plasma is inertial confinement. In this latter case, large scale hydrodynamical instabilities prevent a sufficient large energy deposit and lower the return of the target. Therefore, for both magnetic and inertial confinement technologies, the success of the projects is deeply linked to the theoretical understanding of plasma turbulence and flow instabilities as well as to mathematical and numerical improvements enabling the development of predictive simulation tools.

#### 3.2. Turbulence Modelling

**Participants:** Boniface Nkonga, Richard Pasquetti.

Fluid turbulence has a paradoxical situation in science. The Navier-Stokes equations are an almost perfect model that can be applied to any flow. However, they cannot be solved for any flow of direct practical interest. Turbulent flows involve instability and strong dependence to parameters, chaotic succession of more or less organised phenomena, small and large scales interacting in a complex manner. It is generally necessary to find a compromise between neglecting a huge number of small events and predicting more or less accurately some larger events and trends.

In this direction, CASTOR wishes to contribute to the progress of methods for the prediction of fluid turbulence. Taking benefit of its experience in numerical methods for complex applications, CASTOR works out models for predicting flows around complex obstacles, that can be moved or deformed by the flow, and involving large turbulent structures. Taking into account our ambition to provide also short term methods for industrial problems, we consider methods applying to high Reynolds flows, and in particular, methods hybridizing Large Eddy Simulation (LES) with Reynolds Averaging.

Turbulence is the indirect cause of many other phenomena. Fluid-structure interaction is one of them, and can manifest itself for example in Vortex Induced Motion or Vibration. These phenomena can couple also with liquid-gas interfaces and bring new problems. Of particular interest is also the study of turbulence generated noise. In this field, though acoustic phenomena can also in principle be described by the Navier-Stokes equations, they are not generally numerically solved by flow solvers but rather by specialized linear and nonlinear acoustic solvers. An important question is the investigation of the best way to combine a LES simulation with the acoustic propagation of the waves it produces.

### 3.3. Astrophysical and Environmental flows

**Participants:** Didier Auroux, Hervé Guillard, Boniface Nkonga, Sebastian Minjeaud.

Although it seems inappropriate to address the modeling of experimental devices of the size of a tokamak and for instance, astrophysical systems with the same mathematical and numerical tools, it has long been recognized that the behaviour of these systems have a profound unity. This has for consequence for instance that any large conference on plasma physics includes sessions on astrophysical plasmas as well as sessions on laboratory plasmas. CASTOR does not intend to consider fluid models coming from Astrophysics or Environmental flows for themselves. However, the team is interested in the numerical approximation of some problems in this area as they provide interesting reduced models for more complex phenomena. To be more precise, let us give some concrete examples : The development of Rossby waves <sup>0</sup> a common problem in weather prediction has a counterpart in the development of magnetic shear induced instabilities in tokamaks and the understanding of this latter type of instabilities has been largely improved by the Rossby wave model. A second example is the water bag model of plasma physics that has a lot in common with multi-layer shallow water system.

To give a last example, we can stress that the development of the so-called well-balanced finite volume schemes used nowadays in many domains of mathematical physics or engineering was largely motivated by the desire to suppress some problems appearing in the approximation of the shallow water system.

Our goal is therefore to use astrophysical or geophysical models to investigate some numerical questions in contexts that, in contrast with plasma physics or fluid turbulence, do not require huge three dimensional computations but are still of interest for themselves and not only as toy models.

---

<sup>0</sup>Rosby waves are giant meanders in high altitude wind that have major influence on weather. Oceanic Rossby waves are also know to exist and to affect the world ocean circulation



## 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 models through collaborations with CERE (Centre d'Enseignement et de Recherche en Environnement Atmosphérique, École des Ponts ParisTech): the models from Polyphemus (pollution forecasting from local to regional scales) and Code\_Saturne (urban scale). In regard to other modeling domains, such as meteorology and oceanography, Clime accesses models through co-operation with LOCEAN (Laboratoire d'Océanographie et du climat, UPMC).

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. 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, UK-met, 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 arise through interface conditions. It naturally occurs when the physical conditions are highly different in subdomains of the physical domain in which the flows takes place. Hence interface conditions are intended to describe the exchange (of mass, energy...) between the domains. Again it gives rise to rather unexplored mathematical questions, and for numerics it yields the question of defining a suitable matching at the discrete level, that is requested to preserve the properties of the continuous model.
- By nature the problems we wish to consider involve many different scales (of time or length basically). It raises two families of mathematical questions. In terms of numerical schemes, the multiscale feature induces the presence of stiff terms within the equations, which naturally leads to stability issues. A clear understanding of scale separation helps in designing efficient methods, based on suitable splitting techniques for instance. On the other hand asymptotic arguments can be used to derive hierarchy of models and to identify physical regimes in which a reduced set of equations can be used.

We can distinguish the following fields of expertise

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

## FLUMINANCE Project-Team

### 3. Research Program

#### 3.1. Estimation of fluid characteristic features from images

The measurement of fluid representative features such as vector fields, potential functions or vorticity maps, enables physicists to have better understanding of experimental or geophysical fluid flows. Such measurements date back to one century and more but became an intensive subject of research since the emergence of correlation techniques [29] 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 [42]. Unlike to dense optical flow estimators, the former approach provides techniques that supply only sparse velocity fields. These methods have demonstrated to be robust and to provide accurate measurements for flows seeded with particles. These restrictions and their inherent discrete local nature limit too much their use and prevent any evolutions of these techniques towards the devising of methods supplying physically consistent results and small scale velocity measurements. It does not authorize also the use of scalar images exploited in numerous situations to visualize flows (image showing the diffusion of a scalar such as dye, pollutant, light index refraction, fluorecein,...). At the opposite, variational techniques enable in a well-established mathematical framework to estimate spatially continuous velocity fields, which should allow more properly to go towards the measurement of smaller motion scales. As these methods are defined through PDE's systems they allow quite naturally 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], [10], the setting up of tomographic techniques to carry out 3D velocity measurements [36], 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 [9]. 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 [49], diffusive scalar images, fluid holography [50], 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 [44], [45] or through stochastic filtering approaches [37], [40]. 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 [33] 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 [52], [39]. 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 [30]. Most of the works of the literature indeed comes back to open-loop control approaches [47], [41], [46] or to forcing approaches [38] 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) [35]. 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 [31], [34]. 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 [43]. 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 [32]. This technique, historically embedded in the larger domain of sensor-based control [48], can be properly used to control complex robotic systems or, as we showed it recently, flows [51].

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.

## KALIFFE Project-Team

### 3. Research Program

#### 3.1. Numerical schemes for nonlinear kinetic models in an arbitrary geometry

In this part, we want to focus in the numerical approximation of solutions to kinetic equations (microscopic description) set in a complex geometry with different types of boundary conditions. Many numerical schemes have been proposed to approximate the solutions of nonlinear kinetic equations, but few of them are concerned by the treatment of complex geometry and boundary conditions which have a special interest for applications. In this context, classical structured or unstructured meshes already applied in computational fluid dynamics are not appropriate due to the high dimensional property of kinetic problems. In contrast, the Cartesian mesh makes the numerical method efficient and easy to implement. Indeed, in the framework of the Inria-Calvi project, E. Sonnendrücker and his collaborators have developed several families of methods for solving transport equations in a phase space grid with specific applications to plasma physics. These methods are based on the well known semi-Lagrangian methods. The principle is to solve the equation on a phase space grid, for which the grid points are advected with the flow of the transport equation for a time step and interpolated back periodically on the initial grid. The characteristics can be solved either forward or backward in time leading to the forward semi-Lagrangian or backward semi-Lagrangian schemes. These schemes are particularly well suited for uniform Cartesian grid since they are efficient in term of accuracy (high order scheme), stability (not restricted by a CFL condition) and computational cost (fast to locate the transported grid point).

Our aim is now to use all these techniques in the context of complex geometry and for the treatment of boundary conditions. The difficulty is that obviously grid points are usually not located on the physical boundary when using a Cartesian mesh, thus a suitable numerical method to capture the boundary condition on the complex geometry is required. In order to apply numerical methods previously studied, we want to treat separately the transport equation and the boundary conditions in the complex geometry.

Several numerical methods based on Cartesian mesh have been developed in computational fluid dynamics in last decade. Among these methods, the immersed boundary method (IBM), first introduced by Peskin for the study of biological fluid mechanics problems, has attracted considerable attention because of its use of regular Cartesian grid and great simplification of tedious grid generation task. The basic idea of immersed boundary method is that the effect of the immersed boundary on the surrounding fluid is represented through the introduction of forcing terms in the momentum equations. In conservation laws, two major classes immersed boundary like methods can be distinguished on different discretization types. The first class is Cartesian cut-cell method, which is based on a finite volume method. This conceptually simple approach “cuts” solid bodies out of a background Cartesian mesh. Thus we have several polygons (cut-cells) along the boundary. Then the numerical flux at the boundary of these cut-cells are imposed by using the real boundary conditions. This method satisfies well the conservation laws, however to determine the polygons is still a delicate issue.

Here, we will consider another class of method, based on finite difference method. To achieve a high order interior scheme, several ghost points behind the boundary are added. For instance for solving hyperbolic conservations laws, an inverse Lax-Wendroff type procedure is used to impose some artificial values on the ghost points. The interest of this approach is that it preserves all the flexibility of semi-Lagrangian schemes, that is, high order accuracy, resolution in a uniform Cartesian grid and stability. The members of the project involved in this thematic pole are already studying kinetic and related models and will develop this type of numerical schemes focusing on the following goals:

- **Accuracy.** Achieving arbitrary high accuracy for problems with smooth solutions has been a topic of the utmost importance in the recent years and actively studied by a many researchers and groups worldwide. The project team has been investigating such methods for several years and for various PDE models, for steady and unsteady physical problems, using different formulations among which spectral and semi-Lagrangian methods, discontinuous Galerkin methods and finite volume methods.

In all the cases, we consider numerical methods relying on discretization techniques that best fit to the geometrical characteristics of the problems at hand.

- **Robustness.** On the other hand, these methods should also be capable to accurately describe the underlying physical phenomena that may involve highly variable space and time scales. With reference to this characteristic, several strategies are possible: adaptive local refinement/coarsening of the mesh (*i.e.*  $h$ -adaptivity) and adaptive local variation of the interpolation order (*i.e.*  $p$ -adaptivity). Ideally, these two strategies are combined leading to the so-called  $hp$ -adaptive methods and that will actually represent an ultimate objective of our research activities. Note that both strategies are all local in nature.
- **Efficiency.** Despite the ever increasing performances of microprocessors, the numerical simulation of realistic  $4D$  or  $5D$  kinetic problems is hardly performed on a high-end workstation and parallel computing is a mandatory path. Hence, numerical methods must be adapted to the characteristics of modern parallel computing platforms taking into account their heterogeneity and their hierarchical nature (*e.g.* multiple processors and multiple core systems with complex cache and memory hierarchies, possibly augmented with accelerator cards). Appropriate parallelization strategies need to be designed that combine distributed memory and shared memory paradigms *i.e.* MIMD (multiple instruction, multiple data) and SIMD (single instruction, multiple data) programming models.

### 3.2. Asymptotic Preserving schemes

We develop robust numerical schemes for kinetic equations that also work in the fluid regime. The goal of this part of the project is to propose a new general and systematic strategy that avoids the inversion of the involved time implicit schemes and that allows to apply the microscopic description without any stability constraint on the numerical parameter  $h$ .

#### Development of numerical schemes for stiff problems.

The idea is to combine micro/macro decomposition with penalization techniques for collision operators, leading to completely explicit schemes which are stable in the desired macroscopic limits. The expected schemes should be consistent with the model at both macroscopic and microscopic levels.

However for plasma applications, the Landau-Fokker-Planck operator has a diffusion structure in the velocity variable which induces special behaviors at both mathematical and numerical levels. We will show that the previous methodology can be adapted to overcome the velocity diffusion stiffness in this case. In other words, the obtained numerical schemes are expected to be free of usual diffusion CFL conditions, and will be stable and consistent within the macroscopic limit under consideration. Finally, to decrease the computational cost of the so constructed schemes, which is due to the non local character of the involved collision operator (Boltzmann, Landau, etc), fast computational method for integral operator are needed. On the basis of multi-grid and/or Fast Multi-pole Methods, we shall develop appropriate acceleration methods to our context.

Another important task in this project is to extend the above strategy to the context of a self-consistent Poisson or Maxwell equations. Accurate methods based on finite volume schemes will be developed for high field limit. A stiffness raised by the presence of high electromagnetic fields will also be treated in the same spirit. Such problems are also investigated in the IPSO project (M. Lemou, F. Méhats and N. Crouseilles). Here our strategy is based on a suitable operator decomposition coupled with appropriate IMEX schemes.

**Stability and accuracy issue.** In the framework of Asymptotic-Preserving (AP) schemes, there are few mathematical justifications of stability and uniform accuracy of such approach. A stability analysis has to be performed to rigorously prove that the numerical scheme is stable for small values of small physical parameters even if the time step does not resolve it. This analysis seems to be tricky for fully nonlinear kinetic equations like the Boltzmann equation. Therefore, we focus on simpler models as discrete velocity models (DVM) which have the same properties as the full Boltzmann operator but deal with a finite set of velocity. In this project we are particularly interested by the long time behavior of the numerical solution when it approaches its stationary state. We plan to apply the entropy-entropy dissipation technique to design new numerical approximations. It gives a specific discretization based on finite volume approximation, which allows to control the numerical

entropy production and in some situations, it is often enough to give stability of the numerical solution in the long time asymptotic limit. For general cases, these estimations have to be completed by some discrete functional inequalities.

## 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 [28], [32], [60] are available in one dimensions, let alone in the multidimensional case. We can refer to [58], [35] for some linear dispersive equations, treated with finite-element methods, or to [32] for the first use of advanced high-order compact finite-volume methods for the Serre equations. Recent work undertaken during the ANR MathOCEAN [28] 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 [26]. 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 [29], or undular bores using Boussinesq-type models [39]. Let us also mention [37] for tsunami modeling and [36], [48] 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 [28], [56]. 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 pioneering work of [38] on transparent boundary conditions. Deep studies of these operators have been performed in the case of linear equations, [43], [27], [53]. 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 [49], 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 [40].

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 homogenization technique [25], [24], [52] 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 [25]. Unfortunately, the REV has been shown recently not to exist in urban areas [42]. 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 [34], [41], [44] and applied successfully to urban flood modeling scale experiments [41], [50], [55]. 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 [54] and multiple porosity [42] 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 [33]), Boussinesq-type dispersive models [26]), 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 [47], 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 [46]. 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 [38].

### 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 developed 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 [40]) 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 [30], [31], [34], [41], [42], [44], [50], [54], [55] 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 foreseen that subgrid-scale topographic variations may play a predominant role. Research on the integration of subgrid-scale topography into macrosopic 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 [41], [42], [54], [57].
- 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 [42], [54]. 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 [45], [51]. 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. Inverse Problems**

- **Inverse scattering problems.** The determination of the shape of an obstacle immersed in a fluid medium from some measurements of the scattered field in the presence of incident waves is an important problem in many technologies such as sonar, radar, geophysical exploration, medical imaging and nondestructive testing. Because of its nonlinear and ill-posed character, this inverse obstacle problem (IOP) is very difficult to solve, especially from a numerical viewpoint. The success of the reconstruction depends strongly on the quantity and quality of the measurements, especially on the aperture (range of observation angles) and the level of noise in the data. Moreover, in order to solve IOP, the understanding of the theory for the associated direct scattering problem and the mastery of the corresponding solution methods are fundamental. Magique-3d is involved in the mathematical and numerical analysis of a direct elasto-acoustic scattering problem and of an inverse obstacle scattering problem. More specifically, the purpose of this research axis is to propose a solution methodology for the IOP based on a regularized Newton-type method, known to be robust and efficient.
- **Depth Imaging in the context of DIP.** The challenge of seismic imaging is to obtain an accurate representation of the subsurface from the solution of the full wave equation that is the best mathematical model according to the time reversibility of its solution. The Reverse Time Migration, [71], is a technique for Imaging that is widely used in the industry. It is an iterative process based on the solution of a collection of wave equations. The high complexity of the propagation medium requires the use of advanced numerical methods, which allows one to solve several wave equations quickly and accurately. Magique-3D is involved in Depth Imaging by the way of a collaboration with TOTAL, in the framework of the research program DIP which has been jointly defined by researchers of MAGIQUE-3D and engineers of TOTAL. In this context, MAGIQUE-3D develops new algorithms in order to improve the RTM.

#### **3.2. Modeling**

The main activities of Magique-3D in modeling are the derivation and the analysis of models that are based on mathematical physics and are suggested by geophysical problems. In particular, Magique-3D considers equations of interest for the oil industry and focuses on the development and the analysis of numerical models which are well-adapted to solve quickly and accurately problems set in very large or unbounded domains as it is generally the case in geophysics.

- **High-Order Time Schemes.** Using the full wave equation for migration implies very high computational burdens, in order to get high resolution images. Indeed, to improve the accuracy of the numerical solution, one must significantly reduce the space step, which is the distance between two points of the mesh representing the computational domain. Another solution consists in using high-order finite element methods, which are very accurate even with coarse meshes. However, to take fully advantage of the high-order space discretization, one has to develop also high-order time schemes. The most popular ones for geophysical applications are the modified equation scheme [75], [83] and the ADER scheme [79]. Both rely on the same principle, which consists in applying a Taylor expansion in time to the solution of the wave equation. Then, the high-order derivatives with respect to the time are replaced by high order space operators, using the wave equation. Finally, auxiliary variables are introduced in order to transform the differential equation involving high-order operators into a system of differential equation with low order operators. The advantage of this technique is that it leads to explicit time schemes, which avoids the solution of huge linear systems.

The counterpart is that the schemes are only conditionally stable, which means that the time step is constrained by a CFL (Courant-Friedrichs-Levy) condition. The CFL number defines an upper bound for the time step in such a way that the smaller the space step is, the higher the numbers of iterations will be. Magique-3D is working on the construction and the analysis of new explicit and implicit time schemes which have either larger CFL numbers or local CFL numbers. By this way, the computational costs can be reduced without hampering the accuracy of the numerical solution.

- **Finite Element Methods for the time-harmonic wave equation.** As an alternative to Time-Domain Seismic Imaging, geophysicists are more and more interested by Time-Harmonic Seismic Imaging. The drawback of Time Domain Seismic Imaging is that it requires either to store the solution at each time step of the computation, or to perform many solutions to the wave equation. The advantage of Time Harmonic problems is that the solutions can be computed independently for each frequency and the images are produced with only two computations of the wave equation and without storing the solution. The counterpart is that one has to solve a huge linear system, which can not be achieved today when considering realistic 3D elastic media, even with the tremendous progress of Scientific Computing. Discontinuous Galerkin Methods (DGM), which are well-suited for *hp*-adaptivity, allow the use of coarser meshes without hampering the accuracy of the solution. We are confident that these methods will help us to reduce the size of the linear system to be solved, but they still have to be improved in order to tackle realistic 3D problems. However, there exists many different DGMs, and the choice of the most appropriate one for geophysical applications is still not obvious. Our objectives are **a)** to propose a benchmark in order to test the performances of DGMs for seismic applications and **b)** to improve the most efficient DGMs in order to be able to tackle realistic applications. To these aims, we propose to work in the following directions :
  1. To implement a 2D and 3D solver for time harmonic acoustic and elastodynamic wave equation, based on the Interior Penalty Discontinuous Galerkin Method (IPDGM). The implementation of this solver has started few years ago (see Section 5.1 ) for solving Inverse Scattering Problems and the results we obtained in 2D let us presage that IPDGM will be well-adapted for geophysical problems.
  2. To develop a new hybridizable DG (HDG) [73] for 2D and 3D elastodynamic equation. Instead of solving a linear system involving the degrees of freedom of all volumetric cells of the mesh, the principle of HDG consists in introducing a Lagrange multiplier representing the trace of the numerical solution on each face of the mesh. Hence, it reduces the number of unknowns of the global linear system and the volumetric solution is recovered thanks to a local computation on each element.
  3. To develop upscaling methods for very heterogeneous media. When the heterogeneities are too small compared with the wavelengths of the waves, it is necessary to use such techniques, which are able to reproduce fine scale effects with computations on coarse meshes only.

We also intend to consider finite elements methods where the basis functions are not polynomials, but solutions to the time-harmonic wave equations. We have already developed a numerical method based on plane wave basis functions [78]. The numerical results we have obtained on academic test cases showed that the proposed method is not only more stable than the DGM, but also exhibits a better level of accuracy. These results were obtained by choosing the same plane waves for the basis functions of every element of the mesh. We are now considering a new methodology allowing for the optimization of the angle of incidence of the plane waves at the element level.

Last, we are developing an original numerical methods where the basis functions are fundamental solutions to the Helmholtz equation, such as Bessel or Hankel functions. Moreover, each basis function is not defined element by element but on the whole domain. This allows for reducing the volumetric variational formulation to a surfacic variational formulation.

- **Boundary conditions.** The construction of efficient absorbing boundary conditions (ABC) is very important for solving wave equations. Indeed, wave problems are generally set in unbounded or

very large domains and simulation requires to limit the computational domain by introducing an external boundary, the so-called absorbing boundary. This topic has been a very active research topic during the past twenty years and despite that, efficient ABCs are still to be designed. Classical conditions are constructed to absorb propagating waves and Magique-3D is investigating the way of improving existing ABCs by introducing the modelling of evanescent and glancing waves. For that purpose, we consider the micro-local derivation of the Dirichlet-to-Neumann operator. The interest of our approach is that the derivation does not depend on the geometry of the absorbing surface.

ABCs have been given up when Perfectly Matched Layers (PML) have been designed. PMLs have opened a large number of research directions and they are probably the most routinely used methods for modelling unbounded domains in geophysics. But in some cases, they turn out to be unstable. This is the case for some elastic media. We are thus considering the development of absorbing boundary conditions for elastodynamic media, and in particular for Tilted Transverse Isotropic media, which are of high interest for geophysical applications.

- **Asymptotic modeling.**

During the last 30 years, mathematicians have developed and justified approximate models with multiscale asymptotic analysis to deal with problems involving singularly perturbed geometry or problems with coefficients of different magnitude.

Numerically, all these approximate models are of interest since they allow to mesh the computational domain without taking into account the small characteristic lengths. These techniques lead to a reduction of the computation burden. Unfortunately, these methods do not have penetrated the numerical community since most of the results have been obtained for the two dimensional Laplacian.

The research activity of Magique 3D aims in extending this theory to three-dimensional challenging problems involving wave propagation phenomena. We address time harmonic and time dependent problems for acoustic waves, electromagnetic waves and elastodynamic wave which is a very important topic for industry. Moreover, it remains numerous open questions in the underlying mathematical problems.

Another important issue is the modeling of boundary layers which are not governed by the same model than the rest of the computational domain. It is rather challenging to derive and to justify some matching condition between the boundary layer and the rest of the physical domain for such multiphysical problems.

More precisely, we have worked in 2014 on the following topics:

- Eddy current modeling in the context of electrothermic applications for the design of electromagnetic devices, in collaboration with laboratories Ampère, Laplace, Inria Team MC2, IRMAR, and F.R.S.-FNRS;
- Multiphysics asymptotic modeling of multi perforate plates in turbo reactors in collaboration with Onera.
- Modeling of small heterogeneities for the three dimensional time domain wave equation. This reduced model is a generalization of the so called Lax-Foldy reduced model.
- Modeling the propagation of ultrashort laser pulses in optical fibers.

### 3.3. High Performance methods for solving wave equations

Seismic Imaging of realistic 3D complex elastodynamic media does not only require advanced mathematical methods but also High Performing Computing (HPC) technologies, both from a software and hardware point of view. In the framework of our collaboration with Total, we are optimizing our algorithms, based on Discontinuous Galerkin methods, in the following directions.

- **Minimizing the communications between each processor.** One of the main advantages of Discontinuous Galerkin methods is that most of the calculi can be performed locally on each element of the mesh. The communications are carried out by the computations of fluxes on the faces of the

elements. Hence, there are only communications between elements sharing a common face. This represents a considerable gain compared with Continuous Finite Element methods where the communications have to be done between elements sharing a common degree of freedom. However, the communications can still be minimized by judiciously choosing the quantities to be passed from one element to another.

- **Hybrid MPI and OpenMP parallel programming.** Since the communications are one of the main bottlenecks for the implementation of the Discontinuous Galerkin in an HPC framework, it is necessary to avoid these communications between two processors sharing the same RAM. To achieve this aim, the partition of the mesh is not performed at the core level but at the chip level and the parallelization between two cores of the same chip is done using OpenMP while the parallelization between two cores of two different chips is done using MPI.
- **Porting the code on new architectures.** The goal is to test popular HPC architectures in the context of seismic imaging simulations. Current work concerns the new Intel Many Integrated Core Architecture (Intel MIC) of the Intel Xeon Phi co-processors and the upcoming stand-alone Intel processors.
- **Using Runtimes Systems.** One of the main issue of optimization of parallel code is the portability between different architectures. Indeed, many optimizations performed for a specific architecture are often useless for another architecture. In some cases, they may even reduce the performance of the code. One way to overcome this problem is to use task-based programming models through runtime libraries as StarPU (<http://runtime.bordeaux.inria.fr/StarPU/>) or PaRSEC (<http://icl.cs.utk.edu/parsec/>). However, until now, they have been mostly employed for solving linear algebra problems and our goal is to test their performances on realistic wave propagation simulations. This work is done in the framework of a collaboration with Inria Team Hiepac and Georges Bosilca (University of Tennessee).

We are confident in the fact that the optimizations of the code will allow us to perform large-scale calculations and inversion of geophysical data for models and distributed data volumes with a resolution level impossible to reach in the past.

## MOISE Project-Team

### 3. Research Program

#### 3.1. Introduction

Geophysical flows generally have a number of particularities that make it difficult to model them and that justify the development of specifically adapted mathematical and numerical methods:

- Geophysical flows are non-linear. There is often a strong interaction between the different scales of the flows, and small-scale effects (smaller than mesh size) have to be modelled in the equations.
- Every geophysical episode is unique: a field experiment cannot be reproduced. Therefore the validation of a model has to be carried out in several different situations, and the role of the data in this process is crucial.
- Geophysical fluids are non closed systems, i.e. there are always interactions between the different components of the environment (atmosphere, ocean, continental water, etc.). Boundary terms are thus of prime importance.
- Geophysical flows are often modeled with the goal of providing forecasts. This has several consequences, like the usefulness of providing corresponding error bars or the importance of designing efficient numerical algorithms to perform computations in a limited time.

Given these particularities, the overall objectives of the MOISE project-team described earlier will be addressed mainly by using the mathematical tools presented in the following.

#### 3.2. Numerical Modelling

**Models** allow a global view of the dynamics, consistent in time and space on a wide spectrum of scales. They are based on fluid mechanics equations and are complex since they deal with the irregular shape of domains, and include a number of specific parameterizations (for example, to account for small-scale turbulence, boundary layers, or rheological effects). Another fundamental aspect of geophysical flows is the importance of non-linearities, i.e. the strong interactions between spatial and temporal scales, and the associated cascade of energy, which of course makes their modelling more complicated.

Since the behavior of a geophysical fluid generally depends on its interactions with others (e.g. interactions between ocean, continental water, atmosphere and ice for climate modelling), building a forecasting system often requires **coupling different models**. Several kinds of problems can be encountered, since the models to be coupled may differ in numerous respects: time and space resolution, physics, dimensions. Depending on the problem, different types of methods can be used, which are mainly based on open and absorbing boundary conditions, multi-grid theory, domain decomposition methods, and optimal control methods.

#### 3.3. Data Assimilation and Inverse Methods

Despite their permanent improvement, models are always characterized by an imperfect physics and some poorly known parameters (e.g. initial and boundary conditions). This is why it is important to also have **observations** of natural systems. However, observations provide only a partial (and sometimes very indirect) view of reality, localized in time and space.

Since models and observations taken separately do not allow for a deterministic reconstruction of real geophysical flows, it is necessary to use these heterogeneous but complementary sources of information simultaneously, by using **data assimilation methods**. These tools for **inverse modelling** are based on the mathematical theories of optimal control and stochastic filtering. Their aim is to identify system parameters which are poorly known in order to correct, in an optimal manner, the model trajectory, bringing it closer to the available observations.

**Variational methods** are based on the minimization of a function measuring the discrepancy between a model solution and observations, using optimal control techniques for this purpose. The model inputs are then used as control variables. The Euler Lagrange condition for optimality is satisfied by the solution of the "Optimality System" (OS) that contains the adjoint model obtained by derivation and transposition of the direct model. It is important to point out that this OS contains all the available information: model, data and statistics. The OS can therefore be considered as a generalized model. The adjoint model is a very powerful tool which can also be used for other applications, such as sensitivity studies.

**Stochastic filtering** is the basic tool in the sequential approach to the problem of data assimilation into numerical models, especially in meteorology and oceanography. The (unknown) initial state of the system can be conveniently modeled by a random vector, and the error of the dynamical model can be taken into account by introducing a random noise term. The goal of filtering is to obtain a good approximation of the conditional expectation of the system state (and of its error covariance matrix) given the observed data. These data appear as the realizations of a random process related to the system state and contaminated by an observation noise.

The development of data assimilation methods in the context of geophysical fluids, however, is difficult for several reasons:

- the models are often strongly non-linear, whereas the theories result in optimal solutions only in the context of linear systems;
- the model error statistics are generally poorly known;
- the size of the model state variable is often quite large, which requires dealing with huge covariance matrices and working with very large control spaces;
- data assimilation methods generally increase the computational costs of the models by one or two orders of magnitude.

Such methods are now used operationally (after 15 years of research) in the main meteorological and oceanographic centers, but tremendous development is still needed to improve the quality of the identification, to reduce their cost, and to make them available for other types of applications.

A challenge of particular interest consists in developing methods for assimilating image data. Indeed, images and sequences of images represent a large amount of data which are currently underused in numerical forecast systems. However, despite their huge informative potential, images are only used in a qualitative way by forecasters, mainly because of the lack of an appropriate methodological framework.

### 3.4. Sensitivity Analysis - Quantification of Uncertainties

Due to the strong non-linearity of geophysical systems and to their chaotic behavior, the dependence of their solutions on external parameters is very complex. Understanding the relationship between model parameters and model solutions is a prerequisite to design better models as well as better parameter identification. Moreover, given the present strong development of forecast systems in geophysics, the ability to provide an estimate of the uncertainty of the forecast is of course a major issue. However, the systems under consideration are very complex, and providing such an estimation is very challenging. Several mathematical approaches are possible to address these issues, using either variational or stochastic tools.

**Variational approach.** In the variational framework, the sensitivity is the gradient of a response function with respect to the parameters or the inputs of the model. The adjoint techniques can therefore be used for such a purpose. If sensitivity is sought in the context of a forecasting system assimilating observations, the optimality system must be derived. This leads to the study of second-order properties: spectrum and eigenvectors of the Hessian are important information on system behavior.

**Global stochastic approach.** Using the variational approach to sensitivity leads to efficient computations of complex code derivatives. However, this approach to sensitivity remains local because derivatives are generally computed at specific points. The stochastic approach of uncertainty analysis aims at studying global criteria describing the global variabilities of the phenomena. For example, the Sobol sensitivity index is given by the ratio between the output variance conditionally to one input and the total output variance. The computation of such quantities leads to statistical problems. For example, the sensitivity indices have to be efficiently estimated from a few runs, using semi or non-parametric estimation techniques. The stochastic modeling of the input/output relationship is another solution.



**POMDAPI Project-Team (section vide)**

## **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.

## STEER 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<sup>0</sup>. 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.

---

<sup>0</sup><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<sup>0</sup>, and Dinamica<sup>0</sup>) 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

<sup>0</sup><http://www.clarklabs.org/products/Land-Change-Modeler-Overview.cfm>

<sup>0</sup><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 devising metamodels (or, surrogate models) that approximate an original model and may be simulated, calibrated, or analyzed more efficiently. As for detecting critical data items, this may indicate for which type of data more effort must be spent in the data collection process in order to eventually improve the model's reliability. Finally, SA can be used as one means for validating models, together with validation based on historical data (or, put simply, using training and test data) and validation of model parameters and outputs by experts in the respective application area. 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 magnetic 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. [36]) 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 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. [32].

### 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) [43] for the simulation of core turbulence that has been developed at CEA Cadarache in collaboration with our team [37]. 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 [31].

## 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 MagnetoHydro-Dynamic (MHD) model. It is physically valid only in some parts of the tokamak (at the edges for instance). The fluid model is generally obtained from the hypothesis that the collisions between particles are strong. At Inria, fine collision models are mainly investigated in the KALIFFE team. 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 [38]. 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.

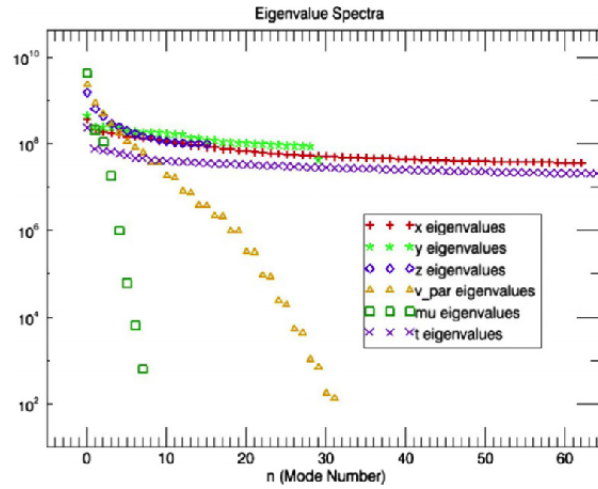


Figure 1. Space and velocity fluctuations spectra (from [38])

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 dependence 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  $\alpha$  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  $\alpha$  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 [41] 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<sup>0</sup>, 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 [42].

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 [34] with the Fourier expansion.

<sup>0</sup>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 intend to write are essentially linear. The nonlinearity is concentrated in a few coupling source terms.

In addition, this method, when written in 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 [44], [30], [40], [39].

## 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 [29] (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 [91]. 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

Cardiac arrhythmias originates from the multiscale organisation of the cardiac action potential from the cellular scale up to the scale of the body. It relates the molecular processes from the cell membranes to the electrocardiogram, an electrical signal on the torso. The spatio-temporal patterns of this propagation is related both to the function of the cellular membrane and of the structural organisation of the cells into tissues, into the organ and final within the body.

Several improvements of current models of the propagation of the action potential will be developed, based on previous work [8] 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 will use high-performance computing techniques in order to explore numerically 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 complex models. We want both to develop simple and fast models of the propagation of cardiac action potentials and improve the solutions to the inverse problems found in cardiac electrical imaging techniques.

The cardiac inverse problem consists in finding the cardiac activation maps or, more generally the whole cardiac electrical activity, from high density body surface electrocardiograms. It is a new and a powerful diagnosis technique, which success would be considered as a breakthrough 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 plan 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 by 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 consider the following strategies:

- using complete propagation models in the inverse problem, like the bidomain equations; for instance in order to localize some electrical sources;
- construct some families of reduced order models, using e.g. statistical learning techniques, which would accurately represent some families of well-identified pathologies;
- construct some 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.

Additionally, we will need to develop numerical techniques dedicated to our simplified eikonal/level-sets 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 need of the medical community. It needs to 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 preconditioning techniques coupled with parallel computing.

## 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 [39] 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 [40] 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 ([45], [47]), and they proved effective in hematopoiesis modelling ([34], [35], [37]). 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 [47]. The nonlocality can be either in the structure variable or in the time variable [34]. 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 [41].



## **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 [16].

#### **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 [6], [7], [8]. 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 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 (<sup>0</sup>, [57], 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 <sup>0</sup> or by ordinary differential equations <sup>0</sup>, or by a mixed representation of Markov processes and ordinary differential equations <sup>0</sup>, 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 <sup>0</sup>. 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 <sup>0</sup> [51] [50]. 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 [59], [60], [58] [40]. Through multiscale analysis, they can be related to stochastic and free boundary models used in cancer modelling.

#### 3.3. 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 [4]) 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.

<sup>0</sup>Drasdo, Hoehme, Block, *J. Stat. Phys.*, 2007

<sup>0</sup>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

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

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

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

<sup>0</sup>B. Perthame, Transport equations in biology, Springer, 2007

### 3.4. Stochastic models

The link between stochastic processes and kinetic equations is a domain already present in our research<sup>0</sup> [53] and that we aim at developing 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 [52], enriching each other. Neuroscience is a domain where this is particularly true because noise contributes significantly to the activity of neurons; this is particularly true for networks where mean field limits are derived from stochastic individual-based models and lead to fundamental questions on well-posedness and behaviours of the system<sup>0</sup>. One strength and originality of our project is our close connections and collaborations not only with probability theorists but also with statisticians.

### 3.5. Agent-based models

Agent-based systems consider each component individually. For example, in agent-based systems of multi-cellular systems the basic modeling unit is the cell, and each cell is considered [54]. This approach has advantages if the population of cells reveals inhomogeneities on small spatial scales as it occurs if organ architecture is represented [57], 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 distinguished by lattice models (e.g. [62]) and lattice-free (or off-lattice) models, in which the position of the cell can change gradually (e.g. [54], [56]). The dynamics of cells in lattice-based models is usually described by rules chosen to mimic the behavior of a cell including its physical behavior. The advantage of this approach is that it is simpler and simulation times for a given number of cells are shorter than in lattice-free models. In contrast, most lattice-free models attempt to parameterize cells by measurable values with a direct physical or biological meaning hence permitting identification of physiologically meaningful parameter ranges. This improves model simulation feasibility, since simulated parameter sensitivity analyses 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 organization processes ([7]). 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 where the interplay of components at many levels is more and more precisely studied [19].

---

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

<sup>0</sup>Cáceres, Carrillo, Perthame *J. Math. Neurosci.* 2011; Pakdaman, Perthame, Salort *Nonlinearity* 2010

## MASAIE Project-Team

### 3. Research Program

#### 3.1. Description

Our conceptual framework is that of Control Theory: the system is described by state variables with inputs (actions on the system) and outputs (the available measurements). Our system is either an epidemiological or immunological system or a harvested fish population. The control theory approach begins with the mathematical modeling of the system. When a “satisfying” model is obtained, this model is studied to understand the system. By “satisfying”, an ambiguous word, we mean validation of the model. This depends on the objectives of the design of the model: explicative model, predictive model, comprehension model, checking hypotheses model. Moreover the process of modeling is not sequential. During elaboration of the model, a mathematical analysis is often done in parallel to describe the behavior of the proposed model. By behavior we intend not only asymptotic behavior but also such properties as observability, identifiability, robustness...

#### 3.2. Structure and modeling

Problems in epidemiology, immunology and virology can be expressed as standard problems in control theory. But interesting new questions do arise. The control theory paradigm, input-output systems built out of simpler components that are interconnected, appears naturally in this context. Decomposing the system into several sub-systems, each of which endowed with certain qualitative properties, allows the behavior of the complete system to be deduced from the behavior of its parts. This paradigm, the toolbox of feedback interconnection of systems, has been used in the so-called theory of large-scale dynamic systems in control theory [23]. Reasons for decomposing are multiple. One reason is conceptual. For example connection of the immune system and the parasitic systems is a natural biological decomposition. Others reasons are for the sake of reducing algorithmic complexities or introducing intended behavior.... In this case subsystems may not have biological interpretation. For example a chain of compartments can be introduced to simulate a continuous delay [19], [20]. Analysis of the structure of epidemiological and immunological systems is vital because of the paucity of data and the dependence of behavior on biological hypotheses. The issue is to identify those parts of models that have most effects on dynamics. The concepts and techniques of interconnection of systems (large-scale systems) is useful in this regard.

In mathematical modeling in epidemiology and immunology, as in most other areas of mathematical modeling, there is always a trade-off between simple models, that omit details and are designed to highlight general qualitative behavior, and detailed models, usually designed for specific situations, including short-terms quantitative predictions. Detailed models are generally difficult to study analytically and hence their usefulness for theoretical purposes is limited, although their strategic value may be high. Simple models can be considered as building blocks of models that include detailed structure. The control theory tools of large-scale systems and interconnections of systems are a mean to conciliate the two approaches, simple models versus detailed systems.

#### 3.3. Dynamic Problems

Many dynamical questions addressed by systems theory are precisely what biologist are asking. One fundamental problem is the problem of equilibria and their stability. To quote J.A. Jacquez

A major project in deterministic modeling of heterogeneous populations is to find conditions for local and global stability and to work out the relations among these stability conditions, the threshold for epidemic take-off, and endemicity, and the basic reproduction number.

The basic reproduction number  $\mathcal{R}_0$  is an important quantity in the study in epidemics. It is defined as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible. The basic reproduction number  $\mathcal{R}_0$  is often considered as the threshold quantity that determines when an infection can invade and persist in a new host population. To the problem of stability is related the problem of robustness, a concept from control theory. In other words how near is the system to an unstable one? Robustness is also in relation with uncertainty of the systems. This is a key point in epidemiological and immunological systems, since there are many sources of uncertainties in these models. The model is uncertain (parameters, functions, structure in some cases), the inputs also are uncertain and the outputs highly variable. That robustness is a fundamental issue and can be seen by means of an example: if policies in public health are to be taken from modeling, they must be based on robust reasons!

### 3.4. Observers

The concept of observer originates in control theory. This is particularly pertinent for epidemiological systems. To an input-output system, is associated the problem of reconstruction of the state. Indeed for a given system, not all the states are known or measured, this is particularly true for biological systems. This fact is due to a lot of reasons: this is not feasible without destroying the system, this is too expensive, there are no available sensors, measures are too noisy...The problem of knowledge of the state at present time is then posed. An observer is another system, whose inputs are the inputs and the outputs of the original system and whose output gives an estimation of the state of the original system at present time. Usually the estimation is required to be exponential. In other words an observer, using the signal information of the original system, reconstructs dynamically the state. More precisely, consider an input-output nonlinear system described by

$$\begin{cases} \dot{x} = f(x, u) \\ y = h(x), \end{cases} \quad (16)$$

where  $x(t) \in \mathbb{R}^n$  is the state of the system at time  $t$ ,  $u(t) \in U \subset \mathbb{R}^m$  is the input and  $y(t) \in \mathbb{R}^q$  is the measurable output of the system.

An observer for the system (1) is a dynamical system

$$\dot{\hat{x}}(t) = g(\hat{x}(t), y(t), u(t)), \quad (17)$$

where the map  $g$  has to be constructed such that: the solutions  $x(t)$  and  $\hat{x}(t)$  of (1) and (2) satisfy for any initial conditions  $x(0)$  and  $\hat{x}(0)$

$$\|x(t) - \hat{x}(t)\| \leq c \|x(0) - \hat{x}(0)\| e^{-at}, \quad \forall t > 0$$

or at least  $\|x(t) - \hat{x}(t)\|$  converges to zero as time goes to infinity.

The problem of observers is completely solved for linear time-invariant systems (LTI). This is a difficult problem for nonlinear systems and is currently an active subject of research. The problem of observation and observers (software sensors) is central in nonlinear control theory. Considerable progress has been made in the last decade, especially by the "French school", which has given important contributions (J.P. Gauthier, H. Hammouri, E. Busvelle, M. Fliess, L. Praly, J.L. Gouzé, O. Bernard, G. Sallet) and is still very active in this area. Now the problem is to identify relevant classes of systems for which reasonable and computable observers can be designed. The concept of observer has been ignored by the modeler community in epidemiology, immunology and virology. To our knowledge one of the first use of an observer in virology was the work of Velasco-Hernandez J., Garcia J. and Kirschner D. [24] in modeling the chemotherapy of HIV, but this observer, based on classical linearization method, is a local observer and does not allow to deal with the nonlinearities.

### **3.5. Delays**

Another crucial issue for biological systems is the question of delays. Delays, in control theory, are traditionally discrete (more exactly, the delays are lags) whereas in biology they usually are continuous and distributed. For example, the entry of a parasite into a cell initiates a cascade of events that ultimately leads to the production of new parasites. Even in a homogeneous population of cells, it is unreasonable to expect that the time to complete all these processes is the same for every cell. If we furthermore consider differences in cell activation state, metabolism, position in the cell cycle, pre-existing stores of nucleotides and other precursors needed for the reproduction of parasites, along with genetic variations in the parasite population, such variations in infection delay times become a near certainty. The rationale for studying continuous delays is supported by such considerations. In the literature on dynamical systems, we find a wealth of theorems dealing with delay differential equations. However they are difficult to apply. Control theory approaches (interconnections of systems) is a mean to study the influence of continuous delays on the stability of such systems. We have obtained some results in this direction [6].

## MODEMIC Project-Team

### 3. Research Program

#### 3.1. Modeling and simulating microbial ecosystems

The chemostat model is quite popular in microbiology and bioprocess engineering [60], [62]. 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:

$$\begin{aligned} \dot{s} &= - \sum_{j=1}^n \frac{1}{y_j} \mu_j(s) x_j + D (s_{in} - s) \\ \dot{x}_i &= \mu_i(s) x_i - D x_i \quad (i = 1 \dots n) \end{aligned} \quad (18)$$

for  $n$  species in concentrations  $x_i$  competing for a substrat 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 [61]. 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 [59].

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 piecewise deterministic Markov processes [58].

### **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 .

## 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  $n$ D, multiscale in space, system of weakly-coupled non conservative transport equations that underlies our approach of gonadal cell dynamics [39],[6]. 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 [7]. 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 [3]. 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 [43] 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  $2nD$  system), control terms operating on the velocities and source term, and formulated from moments of the unknowns, discontinuities both in the velocities and density on internal boundaries of the domain representing the passage from one cell phase to another.

On the theoretical ground, the well-posedness (existence and uniqueness of weak solutions with bounded initial data) has been established in [10], while associated control problems have been studied in the framework of hybrid optimal control [4]. 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 [16] 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 [44], [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 [24].

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 [12], [11] 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 for some models including several scales. We are able to compute the macroscopic speed of propagation and the distribution with respect to the microscopic variable at relevant locations (*e.g.* the edge and the back of the front) in a wide variety of models.

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, age-structured models for epidemiology 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 models. 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, for which we give important contributions too.

These advances are linked to the work of other Inria teams (BANG, 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 populations, and anomalous diffusion.

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.

Emeric Bouin has defended his PhD on December 2nd, 2014. He has accomplished his work under the supervision of Vincent Calvez and Emmanuel Grenier. He has studied propagation phenomena in multiscale models. He has focused on some specific behaviours arising from the multiscale nature of the problem, which are not described by classical reaction-diffusion models. For example, he has discovered unexpected acceleration behaviour in kinetic reaction-transport equations (Bouin, Calvez and Nadin, *Arch. Ration. Mech. Anal.* 2014).

Laetitia Giraldi was a post-doctoral fellow funded by the ANR grant MODPOL under the supervision of Vincent Calvez. She studied thoroughly a biomechanical model for the growth of plant or yeast cells. This new model couples standard equations for the displacement of the cell wall under internal pressure, and a reaction-diffusion equation set on the membrane accounting for the growth pattern has a function of the cell geometry. A rigorous linear stability analysis of the growing spherical shape, together with the development of a stable numerical scheme opens the way to future research in the coupling between growth and geometry.

Alvaro Mateos Gonzalez has started a 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.

### 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.
  - **F. Filbet**, **C. Yang** (Univ. Lyon 1): numerical simulations in 2D in curved geometries.
- 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 subdiffusive processes (just starting)
  - **H. Berry** (Inria BEAGLE), **V. Calvez** (ENS de Lyon, Inria NUMED), **Th. Lepoutre** (Inria DRACULA), **P. Gabriel** (Univ. UVSQ)

This work is also supported by a PEPS project (CNRS) "Physique Théorique et ses Interfaces", led by N. Vauchelet (Univ. Paris 6).

## 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. Vignaux). 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 their 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. cerevisiae*. 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 developed 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 developed 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 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 algorithm 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:

$$\begin{aligned}
 \frac{dC}{dt} &= -KDE \times C \\
 \frac{dP}{dt} &= \lambda_P P \left(1 - \frac{P^{\star\star}}{K}\right) + k_{Q_P} P Q_P - k_{PQ} P - \gamma \times C \times KDE \times P \\
 \frac{dQ}{dt} &= k_{PQ} P - \gamma \times C \times KDE \times Q \\
 \frac{dQ_P}{dt} &= \gamma \times C \times KDE \times Q - k_{Q_P} P Q_P - \delta_{Q_P} Q_P
 \end{aligned} \tag{19}$$

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 [68] in the frame of combustion. It was later used to develop the Kiva code [58] at the Los Alamos National Laboratory, or the Hesione code [63], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [61], 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 [64]. 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<sup>0</sup>. 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.

---

<sup>0</sup><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 [67], and Biot [59] 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 [60]), albeit with *ad hoc* formulations for which compatibility with thermodynamics laws and incompressibility conditions is not established.

### 3.2.3. Tumor and vascularization

The same way the myocardium needs to be perfused for the heart to beat, when it has reached a certain size, tumor tissue needs to be perfused by enough blood to grow. It thus triggers the creation of new blood vessels (angiogenesis) to continue to grow. The interaction of tumor and its micro-environment is an active field of research. One of the challenges is that phenomena (tumor cell proliferation and death, blood vessel adaptation, nutrient transport and diffusion, etc) occur at different scales. A multi-scale approach is thus being developed to tackle this issue. The long term objective is to predict the efficiency of drugs and optimize therapy of cancer.

### 3.2.4. Respiratory tract modeling

We aim at developing a multiscale model of the respiratory tract. Intraparenchymal airways distal from generation 7 of the tracheobronchial tree (TBT), which cannot be visualized by common medical imaging techniques, are modeled either by a single simple model or by a model set according to their order in TBT. The single model is based on straight pipe fully developed flow (Poiseuille flow in steady regimes) with given alveolar pressure at the end of each compartment. It will provide boundary conditions at the bronchial ends of 3D TBT reconstructed from imaging data. The model set includes three serial models. The generation down to the pulmonary lobule will be modeled by reduced basis elements. The lobular airways will be represented by a fractal homogenization approach. The alveoli, which are the gas exchange loci between blood and inhaled air, inflating during inspiration and deflating during expiration, will be described by multiphysics homogenization.

**SISYPHE Project-Team (section vide)**