RESEARCH CENTER

FIELD
Computational Sciences for Biology, Medicine and the Environment

Activity Report 2012

Section Scientific Foundations

Edition: 2013-04-24

ABS Project-Team

3. Scientific Foundations

3.1. Introduction

The research conducted by ABS focuses on two main directions in Computational Structural Biology (CSB), each such direction calling for specific algorithmic developments. These directions are:
- Modeling interfaces and contacts,
- Modeling the flexibility of macro-molecules.

3.2. Modeling Interfaces and Contacts

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

Problems addressed. 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 [1], 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 [46]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [49]. Current investigations follow two routes. From the experimental perspective [31], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [43]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [38].

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 [2], 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 [23], 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 [47], [34]. 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 [39]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

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

[1] 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.

[2] 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 [10]. 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 [24]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [48], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond —a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [30].

A structural alphabet at the atomic level may be seen as an alphabet featuring for an atom of a given type all the conformations this atom may engage into, depending on its neighbors. One way to tackle this problem consists of extending the notions of molecular surfaces used so far, so as to encode multi-body relations between an atom and its neighbors [8]. In order to derive such alphabets, the following two strategies are obvious. On 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.

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 [42]. 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

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 [22]. 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 [21], 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 [20], 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 [20]. 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.

3.3.3. Methodological developments

As outlined by the previous discussion, a number of methodological developments are called for. On the experimental side, the problem of fostering the interpretation of data is under scrutiny. Of particular interest is the disambiguation of proteomics signals (TAP data, mass spectrometry data), and that of density maps coming from electron microscopy. 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. The second one encompasses assessment tools, in order to single out the reconstructions which best comply with the experimental data.

3.4. Modeling the Flexibility of Macro-molecules

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

Problems addressed. 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 and Molecular Dynamics simulations generate ensembles of conformations, called `conformers`. 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 [3]. 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.

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

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [45]. 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 [41], to Morse theory [36] and to analysis of meta-stable states of time series [37] have been proposed.

[3] Assuming local forces are prominent, which in turn subsumes electrostatic interactions are not prominent.

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 [40].

From an algorithmic standpoint, such questions are reminiscent of *shape learning*. Given a collection of samples on an (unknown) *model*, *learning* consists of guessing the model from the samples —the result of this process may be called the *reconstruction*. In doing so, two types of guarantees are sought: topologically speaking, the reconstruction and the model should (ideally!) be isotopic; geometrically speaking, their Hausdorff distance should be small. Motivated by applications in Computer Aided Geometric Design, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years [6]. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison [25]; the study of Morse-like constructions stemming from distance functions to points [33]; the analysis of topological invariants of the model and the samples, and their comparison [26], [27].

Last but not least, gaining insight on such questions would also help to effectively select a reduced set of conformations best representing a larger number of conformations. This selection problem is indeed faced by flexible docking algorithms that need to maintain and/or update collections of conformers for the second stage of the *diffusion - conformer selection - induced fit* complex formation model.

AMIB Project-Team

3. Scientific Foundations

3.1. RNA and protein structures

3.1.1. RNA

Participants: Julie Bernauer, Alain Denise, Rasmus Fonseca, Feng Lou, Yann Ponty, Mireille Régnier, Philippe Rinaudo, Jean-Marc Steyaert.

Common activity with P. Clote (Boston College and Digiteo).

3.1.1.1. From RNA structure to function

We are currently developing a combinatorial approach, based on random generation, to design small and structured RNAs. An application of such a methodology to the Gag-Pol HIV-1 frameshifting site will be carried out with our collaborators at IGM. We hope that, upon capturing the hybridization energy at the design stage, one will be able to gain control over the rate of frameshift and consequently fine-tune the expression of *Gag/Pol*. Our goal is to build these RNA sequences such that their hybridization with existing mRNAs will be favorable to independent folding, and will therefore affect the stability of some secondary structures involved in recoding events. Moreover it has been observed, mainly on bacteria, that some mRNA sequences may adopt alternate folds. Such events are called a conformational switch, or riboswitch. A common feature of recoding events and riboswitches is that some structural element on the mRNA initiates and unusual action of the ribosome, or allows for an alternate fold under some environmental conditions. One challenge is to predict genes that might be subject to riboswitches.

3.1.1.2. Beyond secondary structure

One of our major challenges is to go beyond secondary structure. Over the past decade, few attempts have been made to predict the 3D structure of RNA from sequence only. So far, few groups have taken this leap. Despite the promises shown by their preliminary results, these approaches currently suffer to a limiting scale due to either their high algorithmic complexity or their difficult automation. Using our expertise in algorithmics and modeling, we plan to design original methods, notably within the AMIS-ARN project (ANR BLANC 2008-2012) in collaboration with PRISM at Versailles University and E.Westhof's group at Strasbourg.

1. *Ab initio* modeling: Starting from the predicted RNA secondary structure, we aim to detect *local structural motifs* in it, giving local 3D conformations. We use the resulting partial structure as a flexible scaffold for a multi-scale reconstruction, notably using game theory. We believe the latter paradigm offers a more realistic view of biological processes than global optimization, used by our competitors, and constitutes a real originality of our project.
2. Comparative modeling: we investigate new algorithms for predicting 3D structures by a comparative approach. This involves comparing multiple RNA sequences and structures at a large scale, that is not possible with current algorithms. Successful methods must rely both on new graph algorithms and on biological expertise on sequence-structure relations in RNA molecules.

3.1.1.3. RNA 3D structure evaluation

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. Their function is mainly determined by the structure those molecules adopt as protein and nucleic acids differ from polypeptides and polynucleotides by their spatial organization. This is specially challenging for RNA where structure flexibility is key.

To address those issues, one has to explore the biologically possible spatial configurations of a macromolecule. The two most common techniques currently used in computational structural biology are Molecular Dynamics (MD) and Monte Carlo techniques (MC). Those techniques require the evaluation of a potential or force-field, which for computational biology are often empirical. They mainly consist of a summation of bonded forces associated with chemical bonds, bond angles, and bond dihedrals, and non-bonded forces associated with van der Waals forces and electrostatic charges. Even if there exists implicit solvent models, they are yet not very well performing and still require a lot of computation time.

Our goal, in collaboration with the Levitt lab at Stanford University and H. van den Bedem at the Stanford Synchrotron Radiation Laboratory (Associate Team ITSNAPhttp://pages.saclay.inria.fr/julie.bernauer/EA_ITSNAP/) is to develop knowledge-based (KB) potentials, based on measurements on known RNA 3D structures and provide sampling for experimental structure fitting and docking conformation generation. KB potential are quick to evaluate during a simulation and can be used without having to explicitly address the solvent problem. They can be developed at various levels of representation: -atom, base, nucleotide, domain- and could allow the modelling of a wide size range: from a hairpin to the whole ribosome. We also intend to combine these knowledge-based potentials with other potentials (hybrid modelling) and template-based techniques, allowing accurate modelling and dynamics study of very large RNA molecules. Such studies are still a challenge. We will also study conformations for experimental data fitting by extension the innovative, robotics-inspired Kino-Geometric Sampler conformational search algorithm for proteins to nucleic acids and to include experimental data. KGS models a protein as a kinematic linkage and additionally considers hydrogen bonds that "close" kinematic cycles. In closed kinematic cycles rotatable bonds can no longer be deformed independently without breaking closure. KGS preserves all kinematic cycles, and thus hydrogen bonds, by sampling in a subspace of conformational space defined by all closure constraints. KGS exhibits a singularly large search radius and optimally reduces the number of free parameters. These unique features enable flexible 'docking' of atomic models in the data while moderating the risk of overfitting at low resolution. The KGS procedure can also accommodate knowledge-based potentials to improve evaluation of putative conformations and their interactions (see below).

3.1.2. PROTEINS

Participants: Jérôme Azé, Julie Bernauer, Adrien Guilhot-Gaudeffroy, Jean-Marc Steyaert.

3.1.2.1. Docking and evolutionary algorithms

As mentioned above, the function of many proteins depends on their interaction with one or many partners. Docking is the study of how molecules interact. Despite the improvements due to structural genomics initiatives, the experimental solving of complex structures remains a difficult problem. The prediction of complexes, *docking*, proceeds in two steps: a configuration generation phase or *exploration* and an evaluation phase or *scoring*. As the verification of a predicted conformation is time consuming and very expensive, it is a real challenge to reduce the time dedicated to the analysis of complexes by the biologists. Various algorithms and techniques have been used to perform exploration and scoring [49]. The recent rounds of the CAPRI challenge show that real progress has been made using new techniques [46], [3]. Our group has strong experience in cutting edge geometric modelling and scoring techniques using machine learning strategies for protein-protein complexes. In a collaboration with A. Poupon, INRA-Tours, a method that sorts the various potential conformations by decreasing probability of being real complexes has been developed. It relies on a ranking function that is learnt by an evolutionary algorithm. The learning data are given by a geometric modelling of each conformation obtained by the docking algorithm proposed by the biologists. Objective tests are needed for such predictive approaches. The *Critical Assessment of Predicted Interaction*, CAPRI, a community wide experiment modelled after CASP was set up in 2001 to achieve this goal (http://www.ebi.ac.uk/msd-srv/capri/). First results achieved for CAPRI'02 suggested that it is possible to find good conformations by using geometric information for complexes. This approach has been followed (see section New results). As this new algorithm will produce a huge amount of conformations, an adaptation of the ranking function learning step is needed to handle them. In the near future, we intend to extend our approach to protein-RNA complexes.

Such as in the protein case, the function of RNA molecules also depends on their interaction with one or many partners. Upon interaction, RNA molecules often undergo large conformation changes. Understanding how these molecules interact with proteins would allow better targeting for therapeutic studies. The CAPRI (Critical Assessment of PRediction of Interactions) challenge1 has shown that classical docking procedures largely fail when large conformation change occurs and when RNA is involved. This is especially true for RNA molecules, whose large-scale dynamics remain often unknown. Modeling RNA conformational changes is made hard by the inherent flexible nature of their structure but also by the electrostatics involved. These are hard to model and often lead to computationally expensive simulations. Even if for small RNA molecules, molecular dynamics can be used, such simulations are hard to extend to larger molecules and protein-RNA complexes.

For many diseases, such as cancer and HIV, microRNA molecules play a very important role regulating gene expression by guiding the RISC. Some miRNAs have been shown to suppress tumors and are thus ideal candidates for the development of therapeutic agents. Even if various computational techniques have been developed to predict miRNA targets, none of these consider the structural aspects of the interactions between components of the RISC and miRNA. We aim to target these problems, in collaboration with the Huang lab at HKUST (PHC Procore)

The combination of Voronoi models at a coarse-grained level and powerful machine learning techniques allows the accurate scoring of protein-protein complexes [12]. Our actual machine learning approach for proteins is a combination of several machine learning approaches (evolutionary algorithm, decision trees, decision rules,...). By adapting these approaches to protein-RNA, we would have a fast and efficient technique for scoring large protein-RNA complexes where conformational changes are involved.

Working with RNA instead of protein introduce many major differences in the machine learning approaches. RNA conformations are often smaller than protein conformations, which has an impact on the values of the descriptors used to describe objects. Due to the size differences between RNA and protein, it is often more difficult to generate (during the modeling stage) conformations closed to the biological solution (near native solution). The machine learning algorithms therefore need to take into account all theses specificities to be able to learn good predictive models from data that are not very close to the real solution.

The acquired knowledge on RNA flexibility, dynamics and the importance of the sequence will be a strong advance in the modeling of protein-RNA interactions we are working on. Il will help the development of scoring functions based on Voronoi models for RNA and provide us with the level of flexibility needed in complex conformational search. We also intend to develop hybrid KB potentials for complexes from hybrid RNA KB data. These could be incorporated in leading-edge flexible docking modeling software such as Rosetta.

3.2. Annotations

3.2.1. Word counting

Participants: Alain Denise, Daria Iakovishina, Yann Ponty, Mireille Régnier, Jean-Marc Steyaert.

We aim at enumerating or generating sequences or structures that are *admissible* in the sense that they are likely to possess some given biological property. Team members have a common expertise in enumeration and random generation of combinatorial structures. They have developped computational tools for probability distributions on combinatorial objects, using in particular generating functions and analytic combinatorics. Admissibility criteria can be mainly statistic; they can also rely on the optimisation of some biological parameter, such as an energy function.

The ability to distinguish a significant event from statistical noise is a crucial need in bioinformatics. In a first step, one defines a suitable probabilistic model (null model) that takes into account the relevant biological properties on the structures of interest. A second step is to develop accurate criteria for assessing (or not) their exceptionality. An event observed in biological sequences, is considered as exceptional, and therefore biologically significant, if the probability that it occurs is very small in the null model. Our approach to

compute such a probability consists in an enumeration of good structures or combinatorial objects. Thirdly, it is necessary to design and implement efficient algorithms to compute these formulae or to generate random data sets. Typical examples that motivate research on words and motifs counting are *Transcription Factor Binding Sites*, TFBSs, consensus models of recoding events and some RNA structural motifs. When relevant motifs do not resort to regular languages, one may still take advantage of combinatorial properties to define functions whose study are amenable to our algebraic tools. One may cite secondary structures and recoding events.

Fast development of high throughput technologies has generated a new challenge for computational biology. The main bottlenecks in applications are the computational analysis of experimental data.

As a first example, numerous new assembling algorithms have recently appeared. Still, the comparison of the results arising from these different algorithms led to significant differences for a given genome assembly. Clearly, strong constraints from the underlying technologies, leading to different data (size, confidence,...) are one origin of the problems and a deeper interpretation is needed, in order to improve algorithms and confidence in the results. One objective is to develop a model of errors, including a statistical model, that takes into account the quality of data for the different technologies, and their volume. This is the subject of an international collaboration with V. Makeev's lab (IoGene, Moscow) and MAGNOME project-team. Second, Next Generation Sequencing open the way to the study of structuralvariants in the genome, as recently described in [44]. Defining a probabilistic model that takes into account main dependencies -such as the GC content- is a task o D. Iakovishina's thesis, in a starting 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. Stochastic context-free grammars (SCFG's) have long been used to model both structural and statistical properties of genomic sequences, particularly for predicting the structure of sequences or for searching for motifs. They can also be used to generate random sequences. However, they do not allow the user to fix the length of these sequences. We developed algorithms for random structures generation that respect a given probability distribution on their components. Our approach is based on the concept of *weighted* combinatorial classes, in combination with the so-named *recursive* method for generating combinatorial structures. To that purpose, one first translates the (biological) structures into combinatorial classes, using the *symbolic method*, an algebraic framework developed by Flajolet *et al*. Adding weights to the atoms allows one to bias the probabilities towards the desired distribution. The main issue is to develop efficient algorithms for finding the suitable weights. An implementation was given in the `GenRGenS` software http://www.lri.fr/~genrgens/, and a generic optimizer that automatically derives suitable parameters for a given grammar, is currently being developped.

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [45]: 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 [48], [4], 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 [47].

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 [10], [21].

3.2.3. Programmed -1 ribosomal frameshifting

Participants: Patrick Amar, Jérôme Azé, Alain Denise, Christine Froidevaux, Yann Ponty, Cong Zeng.

During protein synthesis, the ribosome decodes the mRNA by assigning a specific amino acid to each codon or nucleotide triplet. Throughout this process the ribosome moves along the mRNA molecule three nucleotides at a time. However, encounters of specific signals found in mRNA from many viruses lead the ribosome to shift one nucleotide backward thus changing its reading frame. We aim at developing a new computational approach that is able to detect these specific signals in genomic databases in order to better understand the molecular choreography leading to the ribosomal frameshifting, which ultimately will help to rationally design new antiviral drugs. As candidates sequences are expected to be numerous, we aim at developing a ranking method to identify the most relevant sequences. Biological testing of these most promising identified candidates by our collaborators from IGM will help us to refine our computational method.

3.2.4. Knowledge extraction

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

Our main goal is to design semi-automatic methods for annotation. A possible approach is to focus on the way we could discover relevant motifs in order to make more precise links between function and motifs sequence. For instance, a commonly accepted hypothesis is that function depends on the order of the motifs present in a genomic sequence. Likewise we must be able to evaluate the quality of the annotation obtained. This necessitates giving an estimate of the reliability of the results. This may use combinatorial tools described above. It includes a rigorous statement of the validity domain of algorithms and knowledge of the results provenance. We are interested in provenance resulting from workflow management systems that are important in scientific applications for managing large-scale experiments and can be useful to calculate functional annotations. A given workflow may be executed many times, generating huge amounts of information about data produced and consumed. Given the growing availability of this information, there is an increasing interest in mining it to understand the difference in results produced by different executions.

3.2.5. Systems Biology

Participants: Patrick Amar, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Loic Paulevé, Sabine Peres, Mireille Régnier, Jean-Marc Steyaert.

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).

3.2.5.1. Simulations and behavior analysis for metabolism modeling

A great number of methods have been proposed for the study of the behavior of large biological systems. Two methods have been developed and are in use in the team, depending on the specific problems under study : the first one is based on a discrete and direct simulation of the various interactions between the reactants, while the second one deals with an abstract representation by means of differential equations from which we extract various types of features pertaining to the system.

We investigate on the computational modelling step of the cycle by developing a computer simulation system, HSIM, that mimics the interactions of biomolecules in an environment modelling the membranes and compartments found in real cells. In collaboration with biologists from the AMMIS lab. at Rouen we have used HSIM to show the properties of grouping the enzymes of the phosphotransferase system and the glycolytic pathway into metabolons in *E. coli*. In another collaboration with the SYSDIAG Lab (UMR 3145) at Montpellier, we participate at the CompuBioTic project. This is a Synthetic Biology project in the field of medical diagnosis: its goal is to design a small vesicle containing specific proteins 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 human pathologies. We used HSIM to help the design and to test qualitatively and quantitatively this *"biological computer"* before *in vitro*.

Given the set of biochemical reactions which 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 usuall very badly determined. The challenge is therefore to extract information as to the system's behavior while making reasonable asumptions 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. We have developed a software `Mpas` (Metabolic Pathway Analyser Software) that renders the translation in terms of a systems of o.d.e's automatic; then the simulations are also made easy and almost automatic. 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 particular we have been able to prove mathematically the decisive influence of the enzyme PEMT on the Choline/Ethylamine cycles (M. Behzadi's thesis, defended in 2011).

3.2.5.2. Comparison of Metabolic Networks

In the context of a national project, 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 will be studied to find new enzymes allowing the transformation of biomass. The KEGG database contains pathways related to fungi and other species. By analysing these known pathways with rules mining approaches, we would be able to predict new enzymes activities.

3.2.5.3. Signalling networks

AMIB and INRA-BIOS (A. Poupon, Tours) are partners in a two years project ASAM (2011-2012). This project aims to help the understanding of signalling pathways involving G protein-coupled receptors (GPCR) which 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. The aim of ASAM is thus to provide means to semi-automatically construct signalling networks of GPCRS.

BAMBOO Project-Team

3. Scientific Foundations

3.1. Formal methods

The study of symbiosis and of biological interactions more in general is the motivation for the work conducted within BAMBOO, but runs in parallel with another important objective. This concerns to (re)visit classical combinatorial (mainly counting / enumerating) and algorithmic problems on strings and (hyper)graphs, and to explore the new variants / original combinatorial and algorithmic problems that are raised by the main areas of application of this project. As the objectives of these formal methods are motivated by biological questions, they are briefly described together with those questions in the next section.

3.2. 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. Experimentaly, 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 pratically, (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.

3.3. 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 Team

3. Scientific Foundations

3.1. Introduction

As stated above, the research topics of the Beagle Team are centered on the 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 will be 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.We more specifically 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 [47]. 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 [3]. 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 [56]. 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 [4]. 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. To overcome this difficulty, 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 promoterterminator- 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 [64]). As a consequence, this model can be used to study how evolutionary pressures like the ones for robustness or evolvability can shape genome structure [65], [62], [63], [74]. 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 [65], [62], [75]. 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 [55]. 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 recently 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 [49], [48], [50]. 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 [48]. 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 reflexion. 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 [65], and no difference in lifestyles and environment was needed to explain the complexity of the gene regulatory network [48]. 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.

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 [66], [59], 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.

There exist good phylogenic models of punctual mutations on sequences [57], which enable the reconstruction of small parts of ancestral sequences, individual genes for example [67]. But models of whole genome evolution, taking into account large scale events like duplications, insertions, deletions, lateral transfer, rearrangements are just being developped: [77] model punctual mutations as well as duplication and losses of genes, while [52] can reconstruct the evolution of the structure of genomes by inversions. This allows a more comprehensive view of the history of the molecules and the genes, which sometimes have their own historical pattern. But integrative models, considering both nucleotide subsitutions and genome architectures, are still missing.

It is possible to partially reconstruct ancestral genomes for limited cases, by treating separately different types of mutations. It has been done for example for gene content [53], gene order [68], [71], the fate of gene copies after a duplication [61], [45]. All these lead to evolutionary hypotheses on the birth and death of genes [54], on the rearrangements due to duplications [46], [76], on the reasons of variation of genome size [60], [69]. 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 [70], methods for reconstructing the organization of ancestral genomes [72], [51], [73], or for detecting lateral gene transfer events [44], [12]. 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.

BONSAI Project-Team

3. Scientific Foundations

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 [20], [27], [29], [23], [22]. Members of the team have also a strong expertise in text indexing and compressed index data structures [28], [31], [30]. 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 [32], [26], [25], [24], [17] or non-ribosomal peptides [18]. 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.

3.2. High-performance computing

High-performance computing is another tool that we use to achieve our goals. It covers several paradigms: grids, single-instruction, multiple-data (SIMD) instructions or manycore processors such as graphics cards (GPU). For example, libraries like CUDA and OpenCL also facilitate the use of these manycore processors. These hardware architectures bring promising opportunities for time-consuming bottlenecks arising in bioinformatics.

3.3. 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 contexte, Bayesian models and their MCMC sampling allow to approximate probability distributions over parameters and to describe more biologically relevant models.

DYLISS Team

3. Scientific Foundations

3.1. Knowledge representation with constraint programming

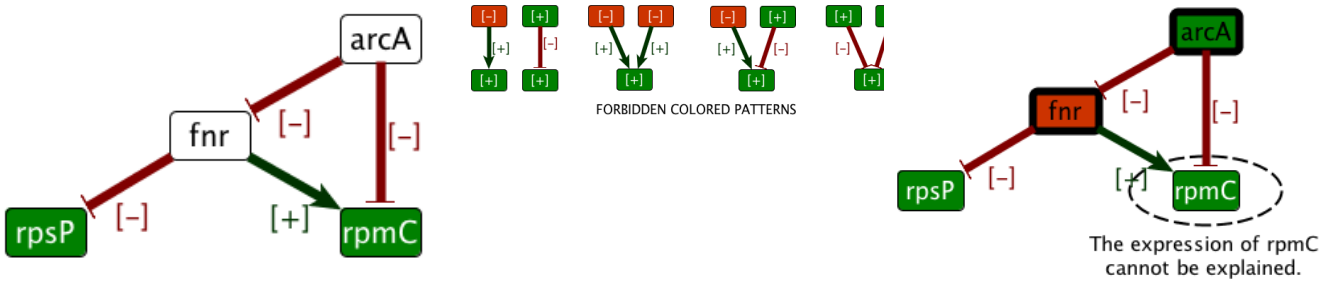
Biological networks are built with data-driven approaches aiming at translating genomic information into a functional map. Most methods are based on a probabilistic framework which defines a probability distribution over the set of models. The reconstructed network is then defined as the most likely 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. Common properties to all solutions are considered as a robust information about the system, as they are independent from the choice of a single solution to the optimization problem[6].

Solving these computational issues requires addressing NP-hard qualitative (non-temporal) issues, based on a notion of causality. We have developed a long-term collaboration with Potsdam University in order to use a logical paradigm named **Answer Set Programming** [27], [30] to solve these optimization issues. Applied on transcriptomic or cancer networks, our methods identified which regions of a large-scale network shall be corrected [1], and proposed robust corrections [5]. 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 optimization issues explored in systems biology.

Using these technologies requires to revisit and reformulate optimization problems at hand in order both to decrease the search space size in the grounding part of the process and to optimize 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.

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 [2]. The system is modeled by an Event Transition Graph, that is, a **Markovian qualitative description of its dynamics** together with quantitative laws which describe the effect of the dynamic transitions over higher scale quantitative measurements. Then, a few time-series quantitative measurements are provided. Following an ergodic assumption and average case analysis properties, we know that a multiplicative accumulation law on a Markov chain asymptotically follows a log-normal law with explicit parameters [29]. This property can be derived into constraints to describe the set of admissible



Step 1 . Regulation knowledge is represented as a signed oriented graph. Edge colors stand for regulatory effects (red/green→ inhibition or activation). Vertex colors stand for gene expression data (red/green→ under or over-expression).

Step 2 . Integrity constraints on the whole colored graph come from the necessity to find a consistent explanation of the link between regulation and expression.

Step 3 . The model allows both the prediction of values (e.g. for fnr in the figure) and the detection of contradictions (e.g. the expression level of rpmC is inconsistant with the regulation in the graph).

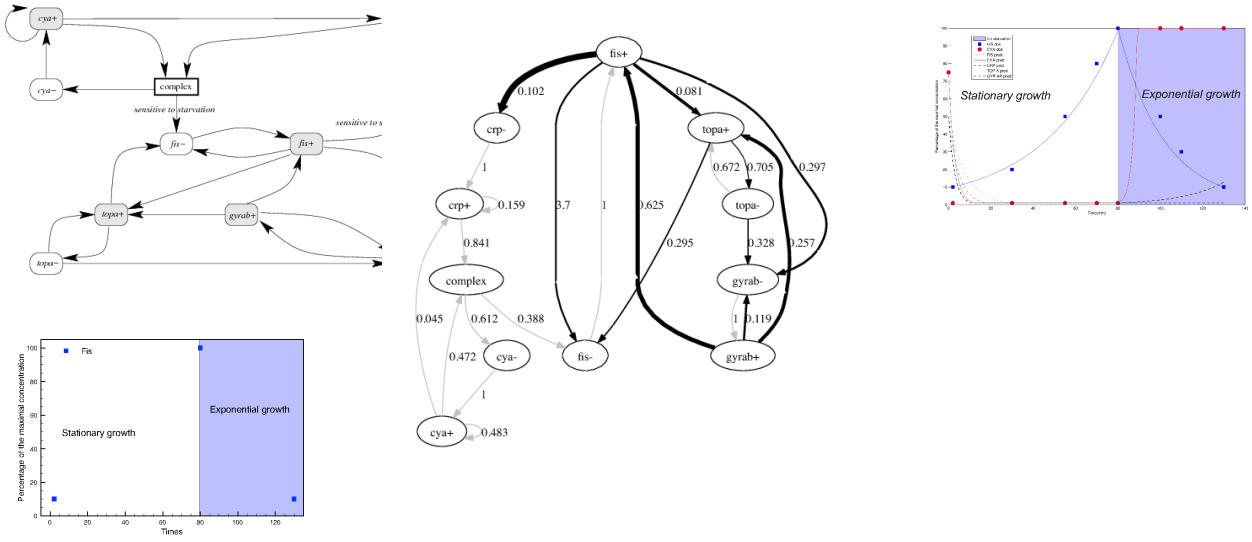
Step 4 . Excerpt from the ASP program.

```
vertex(fnr).                          1{labelV(I ,+;-)}1 :- vertex(I).        receive(I,+) :- labelE(J,I,S), labelV(J,S).
edge(fnr,rpsP).                       labelV(I ,S) :- observedV(I,S).         receive(I,-) :- labelE(J,I,S), labelV(J,T), S≠T.

observedE(fnr,rpsP,-).                1{labelE(J,I,+;-)}1 :- edge(J,I).       :- labelV (I,S), not receive(I,S).
observedV(rpsP,-).                    labelE(J,I,S) :- observedE(J,I,S).
```

Figure 1. Excerpt from the ASP program identifying which expression of non-observed nodes (white nodes) are fixed by partial observations and rules derived from the system dynamics. The 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 [6], [5] .

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. We now plan to apply these techniques to reduced networks representing the main pathways and actors automatically generated from the integrative methods developed in Axis 1. 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.



Input data . Qualitative description of the system dynamics at the transcription level (interaction graph) and 3 concentration measurements of the fis protein (population scale).

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.

Prediction of the Cya protein concentration (red curve) fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in E. Coli response to nutritional stress.

Figure 2. Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems. Identification of key interactions [2].

3.3. Grammatical inference and highly expressive structures

Our main field of expertise in **machine learning** concerns grammatical models with a long -term know-how 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 [4], leading to a tool named Protomata-learner. As an example, this tools allows to properly model the multi-domain function of the protein family TNF, which is impossible with other existing probabilistic-based approach (see Fig. 3). Our future goal is to demonstrate the relevance of formal language theory 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 is order to describe (potentially nested or crossing) long-term dependencies such as contacts of amino-acids that are far in the sequence but close in the 3D protein folding. We wish to extend our expertise towards inferring Context-Free Grammars including the topological information coming from the structural characterization of active sites.

Moving forward to context-free grammars instead of regular patterns increases **parsing** complexity. 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 [31]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [34], [33]. This increases the expressivity of the formalism towards midly 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 [9]. 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.

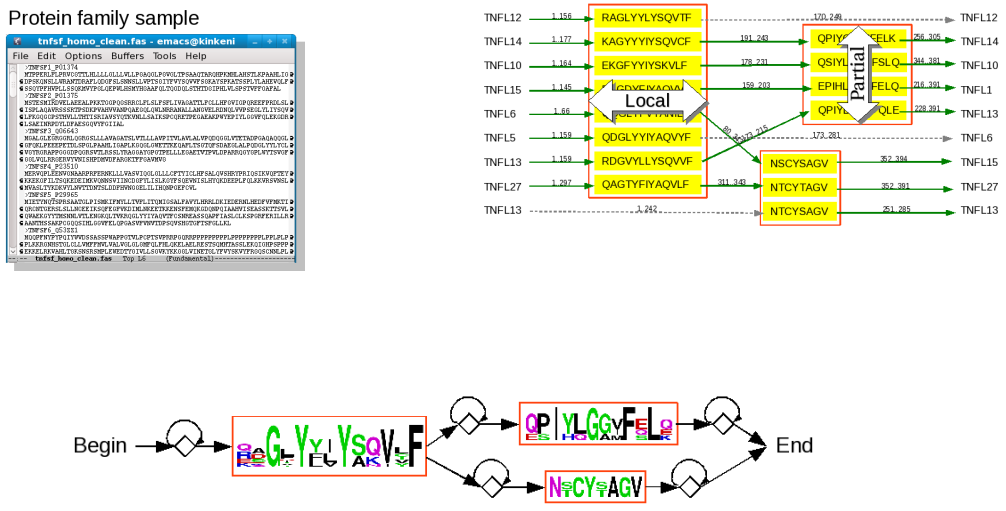


Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences (up left), a partial local alignment is computed (up right) and an automaton is inferred, which models the family and allows to search for its unknown members (down) .*

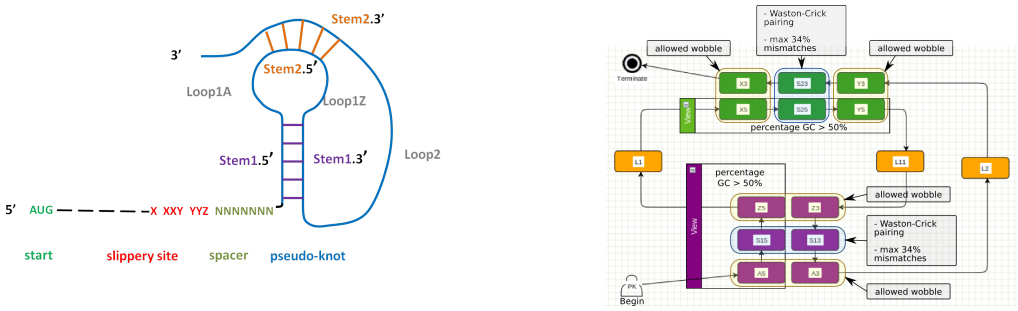


Figure 4. ***Left****: A typical RNA structure: the pseudo-knot.* ***Right****: graphical modeling of a pseudo-knot with String Variable Grammars used in the* ***Logol framework****.*

GENSCALE Team

3. Scientific Foundations

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 fingerprints. 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. Scientific Foundations

3.1. Modeling of bacterial regulatory networks

Participants: Sara Berthoumieux, Eugenio Cinquemani, Johannes Geiselmann, Nils Giordano, Edith Grac, Hidde de Jong, Stéphane Pinhal, Delphine Ropers [Correspondent], Valentin Zulkower.

The adaptation of bacteria to changes in their environment is controlled on the molecular level by large and complex interaction networks involving genes, mRNAs, proteins, and metabolites (Figure 2). The elucidation of the structure of these networks has much progressed as a result of decades of work in genetics, biochemistry, and molecular biology. Most of the time, however, it is not well understood how the response of a bacterium to a particular environmental stress emerges from the interactions between the molecular components of the network. This has called forth an increasing interest in the mathematical modeling of the dynamics of biological networks, in the context of a broader movement called systems biology.

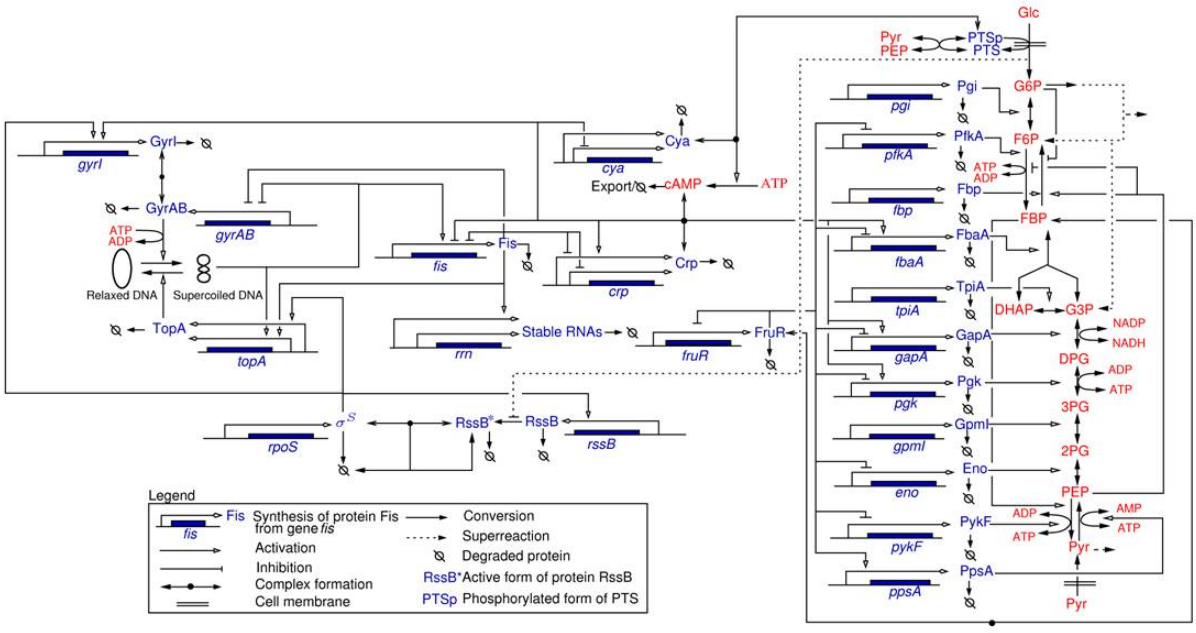


Figure 2. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in E. coli (Baldazzi et al., PLoS Computational Biology, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the 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.

In theory, it is possible to write down mathematical models of biochemical networks, and study these by means of classical analysis and simulation tools. In practice, this is not easy to achieve though, as quantitative data on kinetic parameters are usually absent for most systems of biological interest. Moreover, the models include a large number of variables, are strongly nonlinear and include different time-scales, which make them difficult to handle both mathematically and computationally. A possible approach to this problem has been to use approximate models that preserve essential dynamical properties of the networks. Different approaches have been proposed in the literature, such as the use of approximations of the typical response functions found in gene and metabolic regulation and the reduction of the model dimension by decomposing the system into fast and slow subsystems. These reductions and approximations result in simplified models that are easier to analyze mathematically and for which parameter values can be more reliably estimated from the available experimental data.

Several modeling approaches are exploited in IBIS to gain a better understanding of the ability of *E. coli* to adapt to a various nutritional and other environmental stresses, such as carbon, phosphate, and nitrogen starvation. We are particularly interested in the role of networks of global regulators in shaping the adaptive response of bacteria. Moreover, we study the interactions of these networks with metabolism and the gene expression machinery. These topics involve collaborations with the BEAGLE, COMORE, and CONTRAINTES project-teams at Inria.

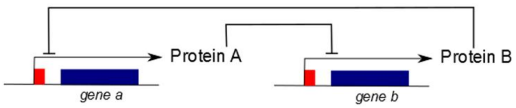
3.2. Analysis, simulation, and identification of bacterial regulatory networks

Participants: Sara Berthoumieux, Eugenio Cinquemani, Johannes Geiselmann, Nils Giordano, Hidde de Jong [Correspondent], Michel Page, François Rechenmann, Delphine Ropers, Diana Stefan, Valentin Zulkower.

Computer simulation is a powerful tool for explaining the capability of bacteria to adapt to sudden changes in their environment in terms of structural features of the underlying regulatory network, such as interlocked positive and negative feedback loops. Moreover, computer simulation allows the prediction of unexpected or otherwise interesting phenomena that call for experimental verification. The use of simplified models of the stress response networks makes simulation easier in two respects. In the first place, model reduction restricts the class of models to a form that is usually easier to treat mathematically, in particular when quantitative information on the model parameters is absent or unreliable. Second, in situations where quantitative precision is necessary, the estimation of parameter values from available experimental data is easier to achieve when using models with a reduced number of parameters.

Over the past few years, we have developed in collaboration with the COMORE project-team a qualitative simulation method adapted to a class of piecewise-linear (PL) differential equation models of gene regulatory networks. The PL models, originally introduced by Leon Glass and Stuart Kauffman, provide a coarse-grained picture of the dynamics of gene regulatory networks. They associate a protein or mRNA concentration variable to each of the genes in the network, and capture the switch-like character of gene regulation by means of step functions that change their value at a threshold concentration of the proteins. The advantage of using PL models is that the qualitative dynamics of the high-dimensional systems are relatively simple to analyze, using inequality constraints on the parameters rather than exact numerical values. The qualitative dynamics of gene regulatory networks can be conveniently analyzed by means of discrete abstractions that transform the PL model into so-called state transition graphs.

The development and analysis of PL models of gene regulatory network has been implemented in the qualitative simulation tool GENETIC NETWORK ANALYZER (GNA) (Section 4.1). GNA has been used for the analysis of several bacterial regulatory networks, such as the initiation of sporulation in *B. subtilis*, quorum sensing in *P. aeruginosa*, the onset of virulence in *E. chrysanthemi*, and environmental biodegradation by *P. putida* mt-2. GNA is currently distributed by the Genostar company, but remains freely available for academic research. The analysis of models of actual bacterial regulatory networks by means of GNA leads to large state transition graphs, which makes manual verification of properties of interest practically infeasible. This has motivated the coupling of GNA to formal verification tools, in particular model checkers that allow properties formulated in temporal logic to be verified on state transition graphs. This has been the subject of collaborations with the POP-ART and VASY project-teams at Inria Grenoble - Rhône-Alpes.



(a)

$$\dot{x}_a = \kappa_a \, s^-(x_a, \theta_a^2) \, s^-(x_b, \theta_b) - \gamma_a \, x_a$$
$$\dot{x}_b = \kappa_b \, s^-(x_a, \theta_a^1) - \gamma_b \, x_b$$
$$s^+(x, \theta) = \begin{cases} 1, & \text{if } x > \theta \\ 0, & \text{if } x < \theta \end{cases}$$
$$s^-(x, \theta) = 1 - s^+(x, \theta)$$

(b)

Figure 3. (a) 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. (b) PLDE model corresponding to the network in (a). Protein A is synthesized at a rate κ_a, if and only if the concentration of protein A is below its threshold θ_a^2 ($x_a < \theta_a^2$) and the concentration of protein B below its threshold θ_b ($x_b < \theta_b$). The degradation of protein A occurs at a rate proportional to the concentration of the protein itself ($\gamma_a \, x_a$).

Recent advances in experimental techniques have led to approaches for measuring cellular processes in real-time on the molecular level, both in single cells and populations of bacteria (Section 3.3). The data sources that are becoming available by means of these techniques contain a wealth of information for the quantification of the interactions in the regulatory networks in the cell. This has stimulated a broadening of the methodological scope of IBIS, from qualitative to quantitative models, and from PL models to nonlinear ODE models and even stochastic models. The group has notably started to work on what is the bottleneck in the practical use of these models, the structural and parametric identification of bacterial regulatory networks from time-series data, in collaboration colleagues from INRA, the University of Pavia (Italy) and ETH Zürich (Switzerland). This raises difficult problems related to identifiability, measurement noise, heterogeneity of data sources, and the design of informative experiments that are becoming increasingly prominent in the systems biology literature.

3.3. High-precision measurements of gene expression in bacteria

Participants: Guillaume Baptist, Sara Berthoumieux, Julien Demol, Johannes Geiselmann [Correspondent], Edith Grac, Jérôme Izard, Hidde de Jong, Stephan Lacour, Yves Markowicz, Corinne Pinel, Stéphane Pinhal, Delphine Ropers, Claire Villiers, Valentin Zulkower.

The aim of a model is to describe the functioning of bacterial regulatory networks so as to gain a better understanding of the molecular mechanisms that control cellular responses and to predict the behavior of the system in new situations. In order to achieve these goals, we have to calibrate the model so that it reproduces available experimental data and confront model predictions with the results of new experiments. This presupposes the availability of high-precision measurements of gene expression and other key processes in the cell.

We have notably resorted to the measurement of fluorescent and luminescent reporter genes, which allow monitoring the expression of a few dozens of regulators in parallel, with the precision and temporal resolution needed for the validation of our models. More specifically, we have constructed transcriptional and translational fusions of key regulatory genes of *E. coli* to fluorescent and luminescent reporter genes (Figure 4). The signals of these reporter genes are measured *in vivo* by an automated, thermostated microplate reader. This makes it possible to monitor in real time the variation in the expression of a few dozens of genes in response to an external perturbation. We have developed an experimental pipeline that resolves most technical difficulties in the generation of reproducible time-series measurements. The pipeline comes with data analysis software that converts the measurements into representations of the time-course of promoter activities that

can be compared with model predictions (Section 4.2). In order to obtain rich information about the network dynamics, we have begun to measure the expression dynamics in both wild-type and mutant cells, using an existing *E. coli* mutant collection. Moreover, we have developed tools for the perturbation of the system, such as expression vectors for the controlled induction of particular genes.

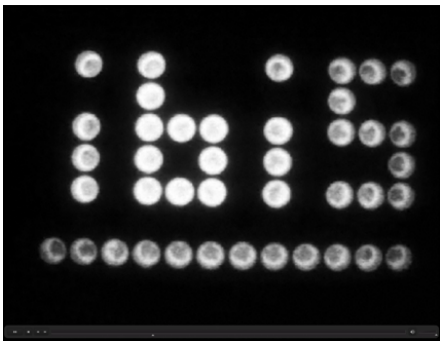


Figure 4. Playful illustration of the principle of reporter genes (see http://ibis.inrialpes.fr for the corresponding movie). A microplate containing a minimal medium (with glucose and acetate) is filmed during 36 hours. Wells contain E. coli bacteria which are transformed with a reporter plasmid containing the luciferase operon (luxCDABE) under control of the acs promoter. This promoter is positively regulated by the CRP-cAMP complex. When bacteria have metabolized all the glucose, the cAMP concentration increases quickly and activates the global regulator CRP which turns on the transcription of the luciferase operon producing the light. The glucose concentration increases from left to right on the microplate, so its consumption takes more time when going up the gradient and the letters appear one after the other. The luciferase protein needs reductive power (FMNH2) to produce light. At the end, when acetate has been depleted, there is no more carbon source in the wells. As a consequence, the reductive power falls and the "bacterial billboard" switches off. Source: Guillaume Baptist.

While reporter gene systems allow the dynamics of gene expression to be measured with high precision and temporal resolution on the level of cell populations, they do not provide information on all variables of interest though. Additional technologies may complement those that we have developed in our laboratory, such as the tools from transcriptomics, proteomics, and metabolomics that are able to quantify the amounts of mRNAs, proteins and metabolites, respectively, in the cells at a given time-point. In addition, for many purposes it is also important to be able to characterize gene expression on the level of single cells instead of cell populations. This requires experimental platforms that measure the expression of reporter genes in isolated cells by means of fluorescence and luminescence microscopy. IBIS has access to these technologies through collaborations with other groups on the local and national level, such as the INSA de Toulouse and the Laboratoire Interdisciplinaire de Physique at the Université Joseph Fourier.

MAGNOME Project-Team

3. Scientific Foundations

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 [20], [6], [10], [11], [56], [57].
- **Integration of heterogenous data**, to build complete knowledge bases for storing and mining information from various sources, and for unambiguously exchanging this information between knowledge bases [1], [4], [41], [44], [29].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [11], [8], [46], [62].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [9], [7], [10].
- **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 [40], [27], [60], [34], [59].

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 [56], [57], [6], [10], [2], [11] and to
- prokaryotes from the lactic bacteria used in winemaking [20], [23], [21], [32], [36], [28].

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 BioRica [26], [40], [60], a framework in which discrete and continuous models can communicate with a clear semantics. Hierarchical models in BioRica can be assembled from existing models, and translated into their execution semantics and then simulated at multiple resolutions through multi-scale stochastic simulation. BioRica 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 [5] [27] and inferring new models through a combination of comparative genomics and successive refinement [51], [52].

MORPHEME Team

3. Scientific Foundations

3.1. Scientific Foundations

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

- **Imaging:** this includes i) definition of the studied populations (experimental conditions) and preparation of samples, ii) definition of relevant quantitative characteristics and optimized acquisition protocol (staining, imaging, $\cdots$) 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 Team

3. Scientific Foundations

3.1. Glossary

WF Optical Wide-Field microscopy.

SDC (Spinning-Disk Confocal microscopy): illumination of the sample with a rotating pattern of several hundred of pinholes for complete simultaneous confocal illumination.

FLIM (Fluorescence Lifetime Microscopy Imaging): imaging of fluorescent molecule lifetimes.

PALM (Photo-Activated Localization Microscopy): high-resolution microscopy using stochastic photo-activation of fluorophores and adjustment of point spread functions [20].

SIM (Structured Illumination Microscopy): high-resolution light microscopy using structured patterns and interference analysis [30].

TIRF (Total Internal Reflectance): 2D optical microscopy using evanescent waves and total reflectance [19].

Cryo-EM (Cryo-Electron Tomography): 3D representation of sub-cellular and molecular objects of 5-20 nanometres, frozen at very low temperatures, from 2D projections using a transmission electron microscope.

3.2. Image restoration for high-resolution microscopy

In order to produce images compatible with the dynamic processes in living cells as seen in video-microscopy, we study the potential of non-local neighborhood filters and image denoising algorithms (e.g. ND-SAFIR software) [6], [2], [7], [4]. The major advantage of these approaches is to acquire images at very low SNR while recovering denoised 2D+T(ime) and 3D+T(ime) images [1]. Such post-acquisition processing can improve the rate of image acquisition by a factor of 100 to 1000 times [5], reducing the sensitivity threshold and allowing imaging for long time regime without cytotoxic effect and photodamages. This approach has been successfully applied to WF, SDC [1], TIRF [19], fast live imaging and 3D-PALM using the OMX system in collaboration with J. Sedat and M. Gustafsson at UCSF [5]. The ND-SAFIR software (see Section 5.1) has been licensed to a large set of laboratories over the world (see Figure 2). 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 [32], PALM [20]) SIM has the strong advantage of versatility when considering the photo-physical properties of the fluorescent probes [30]. 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.

3.3. Dynamic analysis and trajectory computation

3.3.1. Motion analysis and tracking

In time-lapse microscopy, the challenge is to detect and track moving objects. 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 [37]. Data association even combined with sophisticated particle filtering techniques [40] or matching techniques [38] is problematic when tracking several hundreds of similar objects with variable velocities. Developing new optical flow and 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. The goal is to formulate the problem of optical flow estimations in ways that take physical causes of brightness violations into account [26], [31]. In addition, the interpretation of computed flow fields enables to provide spatio-temporal signatures of particular dynamic processes and could help to complete the traffic modelling.

3.3.2. Event detection

Several approaches can be considered for the automatic detection of appearing and vanishing particles (or spots) in WF and TIRF microscopy images. The difficulty is to distinguish motions due to trafficking from the appearing and vanishing spots. Ideally this could be performed by tracking all the vesicles contained in the cell [40], [29]. Among the methods proposed to detect particles in microscopy images [43], [39], none is dedicated to the detection of a small number of particles appearing or disappearing suddenly between two time steps. Our way of handling small blob appearances/dis-appearances originates from the observation that two successive images are redundant and that occlusions correspond to blobs in one image which cannot be reconstructed from the other image [1] (see also [24]).

3.4. Computational 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 [33], [45] aims at modelling systems as a whole in an integrative perspective instead of focusing on independent biophysical processes. One of the goals of these approaches is the cell in silico as investigated at Harvard Medical School (http://vcp.med.harvard.edu/) or the VCell of the University of Connecticut Health Center (http://www.nrcam.uchc.edu/). Previous simulation-based methods have been investigated to explain the spatial organization of microtubules [35] 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) [44] 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 will be to correlate stochastic and dynamical 1D and in silico models studied at the nano-scale in biophysics, to 3D images acquired in vivo at the scale of few hundred nanometres. A difficulty is related to the scale change and statistical aggregation problems (in time and space).

ASCLEPIOS Project-Team

3. Scientific Foundations

3.1. Introduction

Tremendous progress has been made in the automated analysis of biomedical images during the past two decades [91]. Readers who are neophyte to the field of medical imaging will find an interesting presentation of acquisition techniques of the main medical imaging modalities in [82], [80]. Regarding the target applications, a good review of the state of the art can be found in the book *Computer Integrated Surgery* [78], in N. Ayache's article [86] and in the more recent syntheses [87] [91]. The scientific journals *Medical Image Analysis* [73], *Transactions on Medical Imaging* [79], and *Computer Assisted Surgery* [81] 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) [76], [77] or ISBI'2010 (Int. Symp. on Biomedical Imaging) [75].

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 [92], [106]. 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 [109], or to measure some functional properties of the heart from dynamic sequences of Magnetic Resonance [85], Ultrasound or Nuclear Medicine images [93].

Despite these advances and successes, one can notice that statistical models of the anatomy are still very crude, resulting in poor registration results in deformable regions of the body, or between different subjects. If some algorithms exploit the physical modeling of the image acquisition process, only a few actually model the physical or even 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 the observed images and signals, but also more efficient tools to detect anomalies, predict evolutions, simulate and assess 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 the images are multidimensional (3 spatial coordinates and possibly one temporal dimension), but medical protocols tend to include multi-sequence (or multi-parametric) [1] and multi-modal images [2] for each single patient.

[1] 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 acquisition (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 to measure for instance the direction of white matter fibers in the brain (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 subtle higher T2* signal which can be detected with sophisticated image processing techniques.

[2] Multimodal acquisition consists in acquiring on the same patient images from 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 on the physics of image acquisition and observed tissues, as well as on 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 3.4 and 3.5 .

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. Biological Image Analysis

In biology, a huge number of images of living systems are produced every day to study the basic mechanisms of life and pathologies. If some bio-imaging *principles* are the same as the ones used for medical applications (e.g. MR, CT, US, PET or SPECT), the bio-imaging *devices* are usually customized to produce images of higher resolution [3] for the observation of small animals (typically rodents). In addition, Optical Imaging (OI) techniques and biophotonics are developing very fast. This includes traditional or Confocal Microscopy (CM), multi-photon confocal microscopy, Optical Coherent Tomography (OCT), near-infrared imaging, diffuse optical imaging, phased array imaging, etc. A very new and promising development concerns micro-endoscopy, which allows cellular imaging at the end of a very small optical fiber [98].

Most of these imaging techniques can be used for *Molecular Imaging*, an activity aiming at the *in vivo* characterization and measurement of biological processes at cellular and molecular levels. With optical techniques, molecular imaging makes an extensive use of the fluorescent properties of certain molecules (in particular proteins, e.g. GFP [4]) for imaging of gene expression *in vivo*. With other modalities (like PET, SPECT, MR, CT and even US), molecular imaging can use specific contrast agents or radioactive molecules. For clinical applications, the ultimate goal of molecular imaging is to find the ways to probe much earlier the molecular anomalies that are the basis of a disease rather than to image only its end effects [110].

Some of the recent advances made in Medical Image Analysis could be directly applied (or easily adapted) to Biological Image Analysis. However, the specific nature of biological images (higher resolution, different anatomy and functions, different contrast agents, etc.), requires specific image analysis methods (one can refer to the recent tutorial [103] and to the Mouse Brain Atlas Project [84]). This is particularly true when dealing with *in vivo* microscopic images of cells and vessels.

Our research efforts will be focused to the following generic problems applied to *in vivo* microscopic images:

1. quantitative analysis of microscopic images,
2. detection and quantification of variations in temporal sequences,
3. construction of multiscale representations (from micro to macro).

3.4. Computational Anatomy

The objective of Computational Anatomy (CA) is the modeling and analysis of biological variability of the human anatomy. Typical applications cover the simulation of average anatomies and normal variations, the discovery of structural differences between healthy and diseased populations, and the detection and classification of pathologies from structural anomalies [5].

[3] This is the case with micro-MRI, Micro-CT, Micro-US devices, and to a less extent with Micro-SPECT and Micro-PET devices.
[4] Green Fluorescent Protein.
[5] The NIH has lauched 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.

Studying the variability of biological shapes is an old problem (cf. the remarkable book "On Shape and Growth" by D'Arcy Thompson [108]). Significant efforts have been made since that time to develop a theory for statistical shape analysis (one can refer to [90] for a good synthesis, and to the special issue of Neuroimage [107] for recent developments). Despite all these efforts, there is 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 [100]: 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.

3.5. 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 and biology, where CP can be used for instance to better understand the basic processes leading to the apparition 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 [102], [97], [88], [104], [94]). 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 confronting the model with the available biomedical images and signals and possibly also from 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 (e.g. [105], [101]) and the development of new data assimilation methods coping with massive numbers of measurements and unknowns.

There is a hierarchy of modeling levels for CP models of the human body [89]:

- the first level is mainly geometrical, and addresses the construction of a digital description of the anatomy [83], essentially acquired from medical imagery;
- the second level is physical, involving mainly the biomechanical modeling of various tissues, organs, vessels, muscles or bone structures [95];
- the third level is physiological, involving a modeling of the functions of the major biological systems [96] (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 [74].

These different levels of modeling are closely related to each other, and several physiological systems may interact together (e.g. the cardiopulmonary interaction [99]). 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 [96], [97]). 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 biological 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 a 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, ...

3.6. Clinical and Biological 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, this is a necessary condition to see new ideas transformed 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 participate to the stimulation of new ideas and concepts.

ATHENA Project-Team

3. Scientific Foundations

3.1. Computational Diffusion MRI

Diffusion MRI (dMRI) provides a non-invasive way of estimating in-vivo CNS fiber structures using the average random thermal movement (diffusion) of water molecules as a probe. It's a recent field of research with a history of roughly three decades. It was introduced in the mid 80's by Le Bihan et al [59], Merboldt et al [64] and Taylor et al [70]. 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.

Diffusion Tensor Imaging

Shortly after the first acquisitions of diffusion-weighted images (DWI) were made in vivo [65], [66], 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 1-3 mm^3 sized voxel) as an effective self-diffusion tensor given by a 3×3 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 [60], [51]. It has led to many applications in clinical diagnosis of neurological diseases and disorder, neurosciences applications in assessing connectivity of different brain regions, and more recently, therapeutic applications, primarily in neurosurgical planning. An important and very successful application of diffusion MRI has been brain ischemia, following the discovery that water diffusion drops immediately after the onset of an ischemic event, when brain cells undergo swelling through cytotoxic edema.

The increasing clinical importance of diffusion imaging has drived our interest to develop new processing tools for Diffusion 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 [62], [8] and [61]).

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 1mm^3 and 3mm^3 while the physical diameter of fibers can be between 1μm and 30 μm [68], [46]. Research groups currently agree that there is complex fiber architecture in most fiber regions of the brain [67]. 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 [72] such as Q-Ball Imaging (QBI) or Diffusion Spectrum Imaging (DSI) [73], [74], [76] 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 [58] or Diffusion Orientation Transform [79]. 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) [76], [73], [74]. The method requires very strong imaging gradients ($500 \leq b \leq 20000$ 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 [50]. They are at the core of our probabilistic and deterministic tractography algorithms devised to best exploit the full distribution of the fiber ODF (see [48], [3] and [49],[4]).

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 [77], [80] or 4th order Tensor Model [43]. For more details, we refer the reader to our recent article in [53] where we review HOT models and to our article 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.

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 [2], these imaging modalities raise a large amount of mathematical and computational challenges at the core of the research we develop at ATHENA [56].

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 developped 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 [2]. 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. We have also 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 [10], [11]. 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 [10], [11].

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 [55]

3.2. MEG and EEG

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

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

potentials of the pyramidal cells in the cortex. Pyramidal neurons compose approximately 80% of the neurons of the cortex, and it requires at least about 50,000 active such neurons to generate some measurable signal.

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

- First, as acquisition is continuous and is run at a rate up to 1kHz, the amount of data generated by each experiment is huge. Data selection and reduction (finding relevant time blocks or frequency bands) and pre-processing (removing artifacts, enhancing the signal to noise ratio, ...) are largely done manually at present. Making a better and more systematic use of the measurements is an important step to optimally exploit the M/EEG data [1].
- With a proper model of the head and of the sources of brain electromagnetic activity, it is possible to simulate the electrical propagation and reconstruct sources that can explain the measured signal. Proposing better models [6], [9] and means to calibrate them [75] 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.

CORTEX Project-Team

3. Scientific Foundations

3.1. Computational neuroscience

With regards to the progress that has been made in anatomy, neurobiology, physiology, imaging, and behavioral studies, computational neuroscience offers a unique interdisciplinary cooperation between experimental and clinical neuroscientists, physicists, mathematicians and computer scientists. It combines experiments with data analysis and functional models with computer simulation on the basis of strong theoretical concepts and aims at understanding mechanisms that underlie neural processes such as perception, action, learning, memory or cognition.

Today, computational models are able to offer new approaches for the understanding of the complex relations between the structural and the functional level of the brain, thanks to models built at several levels of description. In very precise models, a neuron can be divided in several compartments and its dynamics can be described by a system of differential equations. The spiking neuron approach (*cf.* § 3.2) proposes to define simpler models concentrated on the prediction of the most important events for neurons, the emission of spikes. This allows to compute networks of neurons and to study the neural code with event-driven computations.

Larger neuronal systems are considered when the unit of computation is defined at the level of the population of neurons and when rate coding and/or correlations are supposed to bring enough information. Studying Dynamic Neural Fields (*cf.* § 3.3) consequently lays emphasis on information flows between populations of neurons (feed-forward, feed-back, lateral connectivity) and is well adapted to defining high-level behavioral capabilities related for example to visuomotor coordination.

Furthermore, these computational models and methods have strong implications for other sciences (e.g. computer science, cognitive science, neuroscience) and applications (e.g. robots, cognitive prosthesis) as well (*cf.* § 4.1). In computer science, they promote original modes of distributed computation (*cf.* § 3.5); in cognitive science, they have to be related to current theories of cognition (*cf.* § 3.6); in neuroscience, their predictions have to be related to observed behaviors and measured brain signals (*cf.* § 3.4).

3.2. Computational neuroscience at the microscopic level: spiking neurons and networks

Computational neuroscience is also interested in having more precise and realistic models of the neuron and especially of its dynamics. We consider that the latter aspect cannot be treated at the single unit level only; it is also necessary to consider interactions between neurons at the microscopic scale.

On one hand, compartmental models describe the neuron at the inner scale, through various compartments (axon, synapse, cellular body) and coupled differential equations, allowing to numerically predict the neural activity at a high degree of accuracy. This, however, is intractable if analytic properties are to be derived, or if neural assemblies are considered. We thus focus on phenomenological punctual models of spiking neurons, in order to capture the dynamic behavior of the neuron isolated or inside a network. Generalized conductance based leaky integrate and fire neurons (emitting action potential, i.e. spike, from input integration) or simplified instantiations are considered in our group.

On the other hand, one central issue is to better understand the precise nature of the neural code. From rate coding (the classical assumption that information is mainly conveyed by the firing frequency of neurons) to less explored assumptions such as high-order statistics, time coding (the idea that information is encoded in the firing time of neurons) or synchronization aspects. At the biological level, a fundamental example is the synchronization of neural activities, which seems to play a role in, e.g., olfactory perception: it has been observed that abolishing synchronization suppresses the odor discrimination capability. At the computational

level, recent theoretical results show that the neural code is embedded in periodic firing patterns, while, more generally, we focus on tractable mathematical analysis methods coming from the theory of nonlinear dynamical systems.

For both biological simulations and computer science emerging paradigms, the rigorous simulation of large neural assemblies is a central issue. Our group is at the origin, up to our best knowledge, of the most efficient event-based neural network simulator (Mvaspike), based on well-founded discrete event dynamic systems theory, and now extended to other simulation paradigms, thus offering the capability to push the state of the art on this topic.

3.3. Computational neuroscience at the mesoscopic level: dynamic neural field

Our research activities in the domain of computational neurosciences are also interested in the understanding of higher brain functions using both computational models and robotics. These models are grounded on a computational paradigm that is directly inspired by several brain studies converging on a distributed, asynchronous, numerical and adaptive processing of information and the continuum neural field theory (CNFT) provides the theoretical framework to design models of population of neurons.

This mesoscopic approach underlines the fact that the number of neurons is very high, even in a small part of tissue, and proposes to study neuronal models in a continuum limit where space is continuous and main variables correspond to synaptic activity or firing rates in population of neurons. This formalism is particularly interesting because the dynamic behavior of a large piece of neuronal tissue can be studied with differential equations that can integrate spatial (lateral connectivity) and temporal (speed of propagation) characteristics and display such interesting behavior as pattern formation, travelling waves, bumps, etc.

The main cognitive tasks we are currently interested in are related to sensorimotor systems in interaction with the environment (perception, coordination, planning). The corresponding neuronal structures we are modeling are part of the cortex (perceptive, associative, frontal maps) and the limbic system (hippocampus, amygdala, basal ganglia). Corresponding models of these neuronal structures are defined at the level of the population of neurons and functioning and learning rules are built from neuroscience data to emulate the corresponding information processing (filtering in perceptive maps, multimodal association in associative maps, temporal organization of behavior in frontal maps, episodic memory in hippocampus, emotional conditioning in amygdala, selection of action in basal ganglia). Our aim is to iteratively refine these models, implement them on autonomous robots and make them cooperate and exchange information, toward a completely adaptive, integrated and autonomous behavior.

3.4. Brain Signal Processing

The observation of brain activity and its analysis with appropriate data analysis techniques allow to extract properties of underlying neural activity and to better understand high level functions. This study needs to investigate and integrate, in a single trial, information spread in several cortical areas and available at different scales (MUA, LFP, ECoG, EEG).

One major problem is how to be able to deal with the variability between trials. Thus, it is necessary to develop robust techniques based on stable features. Specific modeling techniques should be able to extract features investigating the time domain and the frequency domain. In the time domain, template-based unsupervised models allows to extract graphic-elements. Both the average technique to obtain the templates and the distance used to match the signal with the templates are important, even when the signal has a strong distorted shape. The study of spike synchrony is also an important challenge. In the frequency domain, features such as phases, frequency bands and amplitudes contain different pieces of information that should be properly identified using variable selection techniques. In both cases, compression techniques such as PCA or ICA can reduce the fluctuations of the cortical signal. Then, the designed models have to be able to track the dynamic evolution of these features over the time.

Another problem is how to integrate information spreading in different areas and relate this information in a proper time window of synchronization to behavior. For example, feedbacks are known to be very important to better understand the closed-loop control of a hand grasping movement. However, from the preparatory signal and the execution of the movement to the visual and somatosensory feedbacks, there is a delay. It is thus necessary to use stable features to build a mapping between areas using supervised models taking into account a time window shift.

Several recoding techniques are taken into account, providing different kinds of information. Some of them provide very local information such as multiunit activities (MUA) and local field potential (LFP) in one or several well-chosen cortical areas. Other ones provide global information about close regions such as electrocorticography (ECoG) or the whole scalp such as electroencephalography (EEG). If surface electrodes allow to easily obtain brain imaging, it is more and more necessary to better investigate the neural code.

3.5. Connectionist parallelism

Connectionist models, such as neural networks, are among the first models of parallel computing. Artificial neural networks now stand as a possible alternative with respect to the standard computing model of current computers. The computing power of these connectionist models is based on their distributed properties: a very fine-grain massive parallelism with densely interconnected computation units.

The connectionist paradigm is the foundation of the robust, adaptive, embeddable and autonomous processings that we aim at developing in our team. Therefore their specific massive parallelism has to be fully exploited. Furthermore, we use this intrinsic parallelism as a guideline to develop new models and algorithms for which parallel implementations are naturally made easier.

Our approach claims that the parallelism of connectionist models makes them able to deal with strong implementation and application constraints. This claim is based on both theoretical and practical properties of neural networks. It is related to a very fine parallelism grain that fits parallel hardware devices, as well as to the emergence of very large reconfigurable systems that become able to handle both adaptability and massive parallelism of neural networks. More particularly, digital reconfigurable circuits (e.g. FPGA, Field Programmable Gate Arrays) stand as the most suitable and flexible device for low cost fully parallel implementations of neural models, according to numerous recent studies in the connectionist community. We carry out various arithmetical and topological studies that are required by the implementation of several neural models onto FPGAs, as well as the definition of hardware-targetted neural models of parallel computation.

This research field has evolved within our team by merging with our activities in behavioral computational neuroscience. Taking advantage of the ability of the neural paradigm to cope with strong constraints, as well as taking advantage of the highly complex cognitive tasks that our behavioral models may perform, a new research line has emerged that aims at defining a specific kind of brain-inspired hardware based on modular and extensive resources that are capable of self-organization and self-recruitment through learning when they are assembled within a perception-action loop.

3.6. The embodiment of cognition

Recent theories from cognitive science stress that human cognition emerges from the interactions of the body with the surrounding world. Through motor actions, the body can orient toward objects to better perceive and analyze them. The analysis is performed on the basis of physical measurements and more or less elaborated emotional reactions of the body, generated by the stimuli. This elicits other orientation activities of the body (approach and grasping or avoidance). This elementary behavior is made possible by the capacity, at the cerebral level, to coordinate the perceptive representation of the outer world (including the perception of the body itself) with the behavioral repertoire that it generates either on the physical body (external actions) or on a more internal aspect (emotions, motivations, decisions). In both cases, this capacity of coordination is acquired from experience and interaction with the environment.

The theory of the situatedness of cognition proposes to minimize representational contents (opposite to complex and hierarchical representations) and privileges simple strategies, more directly coupling perception and action and more efficient to react quickly in the changing environment.

A key aspect of this theory of intelligence is the Gibsonian notion of affordance: perception is not a passive process and, depending on the current task, objects are discriminated as possible "tools" that could be used to interact and act in the environment. Whereas a scene full of details can be memorized in very different and costly ways, a task-dependent description is a very economical way that implies minimal storage requirements. Hence, remembering becomes a constructive process.

For example with such a strategy, the organism can keep track of relevant visual targets in the environment by only storing the movement of the eye necessary to foveate them. We do not memorize details of the objects but we know which eye movement to perform to get them: The world itself is considered as an external memory.

Our agreement to this theory has several implications for our methodology of work. In this view, learning emerges from sensorimotor loops and a real body interacting with a real environment are important characteristics for a learning protocol. Also, in this view, the quality of memory (a flexible representation) is preferred to the quantity of memory.

DEMAR Project-Team

3. Scientific Foundations

3.1. Modelling and identification of the sensory-motor system

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

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 contribution concern the following aspects:

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

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

3.2. Synthesis and Control of Human Functions

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

We aim at developping realistic solutions for real clinical problems expressed by patients and medical staff. Different approaches and specifications are developped to answer to those issues in short, mid or long terms. This research axis is therefore obviously strongly related to clinical application objectives. Even though applications can appear very different, the problematic and constraints are usually similar in the context of electrical stimulation: classical desired trajectory tracking is not possible, robustness to disturbances is critical, possible observations of system are limited. Furthermore there is an interaction between body segments under voluntary control of the patient and body segments under artificial control. Finally, this axis relies on modelling and identification results obtained in the first axis and on the technological solutions and approaches developped in the third axis (Neuroprostheses). The robotics framework involved in DEMAR work is close to the tools used and developped by BIPOP team in the context of bipedal robotics. There is no national teams 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, Guy Cathébras, Fabien Soulier, Serge Bernard.

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

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

Keeping the electronics close to the electrode appears to be a good, if not the unique, solution for a complex FES system; this is the concept according to which we direct our neuroprosthesis design and development, in close relationship with other objectives of our project-team (control for instance) but also in close collaboration with medical and industrial partners.
Our efforts are mainly directed to implanted FES system 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 Team

3. Scientific Foundations

3.1. Structured coupled low- and high-level visual perception

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

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

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 .

A driving force behind research in GALEN has been the understanding that these three aspects are tightly coupled. In particular, efficient optimization can be achieved by resorting to sparse image representations that 'shortlist' putative solutions and/or by working with scoring functions that can be efficiently optimized. However, the accuracy of a scoring function is largely affected by the breadth of relationships that it accommodates, as well as the completeness of the employed image representation. Determining the tradeoff between these two requirements is far from obvious and often requires approaches customized to the particular problem setting addressed. Summarizing, even though the three problems outlined above can be addressed in isolation, an integrated end-to-end approach is clearly preferable, both for computational efficiency and performance considerations.

Research in GALEN has therefore dealt with the following problem aspects: first, developing a generic and reliable low-level image representation that can be used transversally across multiple tasks. The use of learning-based techniques has been pursued for boundary detection and symmetry detection in [32], yielding state-of-the-art results, while in [27] trajectory grouping was used to come up with a mid-level representation of spatio-temporal data. Complementary to the detection of geometric structures, we have also explored methods for their description both for image and surface data [17]. We are currently pursuing the formulation of the task in structured prediction terms, which will hopefully allow us to exploit the geometrical interdependencies among symmetry and boundary responses.

Second, we have worked on learning scoring functions for detection with deformable models that can leverage upon the developed low-level representations, while also being amenable to efficient optimization. Building on our earlier work on using boundary and symmetry detector responses to perform groupwise registration within categories we used discriminative learning to train hierarchical object models that rely on shape-based representations; these were successfully applied to the detection of shape-based categories, while we are currently pursuing their integration with appearance-based models.

Third, efficient optimization for deformable models was pursued in [18], where we have developed novel techniques for object detection that employ combinatorial optimization tools (A* and Branch-and-Bound) to tame the combinatorial complexity; in particular our work has a best-case performance that is logarithmic in the number of pixels, while our work in [18] allows us to further accelerate object detection by integrating low-level processing (convolutions) with a bounding-based object detection algorithm. Working on a different approach, in [10] we have pursued the exploitation of reinforcement-learning to optimize over the set of shapes derivable from shape grammars. We are currently pursuing a full-fledged bounding-based inference algorithm, which will integrate the tasks of boundary detection and grouping in a single, integrated object detection algorithm.

3.2. Machine Learning & Structure Prediction

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

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

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

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

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

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:

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

where Ω 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:

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

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:

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

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

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

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

3.4. Discrete Biomedical Image Perception

A wide variety of tasks in medical image analysis can be formulated as discrete labeling problems. In very simple terms, a discrete optimization problem can be stated as follows: we are given a discrete set of variables $\mathcal{V}$, all of which are vertices in a graph $\mathcal{G}$. The edges of this graph (denoted by $\mathcal{E}$) encode the variables' relationships. We are also given as input a discrete set of labels $\mathcal{L}$. We must then assign one label from $\mathcal{L}$ to each variable in $\mathcal{V}$. However, each time we choose to assign a label, say, x_{p_1} to a variable p_1, we are forced to pay a price according to the so-called *singleton* potential function $g_p(x_p)$, while each time we choose to assign a pair of labels, say, x_{p_1} and x_{p_2} to two interrelated variables p_1 and p_2 (two nodes that are connected by an edge in the graph $\mathcal{G}$), we are also forced to pay another price, which is now determined by the so called *pairwise* potential function $f_{p_1p_2}(x_{p_1}, x_{p_2})$. Both the singleton and pairwise potential functions are problem specific and are thus assumed to be provided as input.

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

$$\arg\min_{\{x_p\}} \mathcal{P}(g,f) = \sum_{p\in\mathcal{V}} g_p(x_p) + \sum_{(p_1,p_2)\in\mathcal{E}} f_{p_1p_2}(x_{p_1}, x_{p_2}). \tag{8}$$

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

$$\arg\min_{\{x_p\}} \mathcal{P}(g,f) = \sum_{p\in\mathcal{V}} g_p(x_p) + \sum_{(p_1,p_2)\in\mathcal{E}} f_{p_1p_2}(x_{p_1}, x_{p_2}) + \sum_{C_i\in\mathcal{E}} f_{p_1\cdots p_n}(x_{p_{i^1}}, \cdots, p_{x_{i|C_i|}}). \tag{9}$$

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

MNEMOSYNE Team

3. Scientific Foundations

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 controling 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 [31]. 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 [27], most of these frameworks (e.g. Soar, ACT-R), though sometimes justified on biological grounds, do not go up to a *connectionist* neuronal implementation. Furthermore, because of the complexity of the resulting frameworks, they are restricted to simple symbolic interfaces with the internal and external world and to (relatively) small-sized internal structures. Our main research objective is undoubtly to build such a general purpose cognitive architecture (to model the brain *as a whole* in a systemic way), using a connectionist implementation and able to cope with a realistic environment.

3.2. Computational Neuroscience

From a general point of view, computational neuroscience can be defined as the development of methods from computer science and applied mathematics, to explore more technically and theoretically the relations between structures and functions in the brain [33], [21]. 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 specific structures, we aim at exploring synergies between structures. Consequently, we target interfaces and interplay between heterogenous modes of computing, which is rarely addressed in classical computational neuroscience.

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

Here, computational neuroscience means mimicking original computations made by the neuronal substratum and mastering their corresponding properties: computations are distributed and adaptive; they are performed without an homonculus or any central clock. Numerical schemes developped for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [18], [26] and were the basis for several contributions in our group [32], [29], [34]. 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 [19]; the associated formalism [24] 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 developped for dynamic neural fields [16], that demonstrated a richness of behavior [20] 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 [29], [30] and their use for observing the emergence of typical cortical properties [23]. 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 [25] 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 for 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 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 [22], [15]. 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 skills that have been learned previously. That is the reason why cognitive architectures are often referred to as systems of memory, communicating and sharing information for problem solving. Instead of the classical view in machine learning of a flat architecture, a more complex network of modules must be considered here, as it is the case in the domain of deep learning. In addition, our systemic approach brings the question of incrementally building such a system, with a clear inspiration from developmental sciences. In this perspective, modules can generate internal signals corresponding to internal goals, predictions, error signals, able to supervise the learning of other modules (possibly endowed with a different learning rule), supposed to become autonomous after an instructing period. A typical example is that of episodic learning (in the hippocampus), storing declarative memory about a collection of past episods and supervising the training of a procedural memory in the cortex.

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

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

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

3.4. Autonomous Robotics

Autonomous robots are not only convenient platforms to implement our algorithms; the choice of such platforms is also motivated by theories in cognitive science and neuroscience indicating that cognition emerges from interactions of the body in direct loops with the world 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 [17], 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. Scientific Foundations

3.1. Neural networks dynamics

The study of neural networks is certainly motivated by the dream 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 conductance-based Integrate and Fire models [2], [64], [65], models of epilepsy [77], effects of synaptic plasticity (see the corresponding section below).
Selected publications on this topic.

3.2. Mean-field approaches

Modelling neural activity at scales integrating the effect of thousands of neurons is of central importance for several reasons. First, most imaging techniques are not able to measure individual neuron activity ("microscopic" scale), but are instead measuring mesoscopic effects resulting from the activity of several hundreds to several hundreds of thousands of neurons. Second, anatomical data recorded in the cortex reveal the existence of structures, such as the cortical columns, with a diameter of about 50μm to 1mm, containing of the order of one hundred to one hundred thousand neurons belonging to a few different species. The description of this collective dynamics requires models which are different from individual neurons models. In particular, when the number of neurons is large enough averaging effects appear, and the collective dynamics is well described by an effective mean-field, summarizing the effect of the interactions of a neuron with the other neurons, and depending on a few effective control parameters. This vision, inherited from statistical physics requires that the space scale be large enough to include a large number of microscopic components (here neurons) and small enough so that the region considered is homogeneous.
Our group is developing mathematical and numerical methods allowing on one hand to produce dynamic mean-field equations from the physiological characteristics of neural structure (neurons type, synapse type and anatomical connectivity between neurons populations), and on the other so simulate those equations. 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. Jansen-Rit, 95) 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 the DYVA team at INT, 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 integrodifferential 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 different time scales and of noise.
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 hidden probability 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) [3],[13], [63] . 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 [43], [12]. On the other hand we are proposing new algorithms for data processing [22]. We have developed a C++ software for spike train statistics based on Gibbs distributions analysis and freely available at http://enas.gforge.inria.fr/v3/. We are using this software in collaboration with several biologist groups involved in the analysis of retina spike trains (Centro de Neurociencia Valparaiso; Molecular Biology Lab, Princeton; Institut de la vision, Paris) [26], [22], [55].
Selected publications on this topic.

3.5. Synaptic Plasticity

Neural networks show amazing abilities for information storage and processing, and stimulus-dependent activity shaping, to evolve and adapt. 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 behavioural states (e.g. transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge. On experimental grounds, different synaptic plasticity mechanisms have been exhibited from the Hebbian's ones (Hebb, 1949) to Long Term Potentiation (LTP) and Long Term Depression (LTD), and more recently to Spike Time Dependent Plasticity (STDP) (Markram, Lubke, Frotscher, and Sakmann, 1997; Bi and Poo, 2001). Synaptic plasticity implies that activity guides the way synapses evolve; but the resulting connectivity structure in turn can raise new dynamical regimes. This interaction becomes even more complex if the considered basic architecture is not feed-forward but includes recurrent synaptic links, like in cortical structures. Understanding this mutual coupling between dynamics and topology and its effects on the computations made by the network is a key problem in computational neuroscience.

Our group is developing mathematical and numerical methods to analyse this mutual interaction. Especially, 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 [73], [74],[4]. Also, we are currently studying the conjugated effects of synaptic and intrinsic plasticity in collaboration with H. Berry (Inria Beagle) and B. Delord, J. Naudé, ISIR team, Paris.
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 mainly focused on the study of retinal processing and motion integration at the level of V1 and MT cortical areas.

At the retina level, we are modelling its circuitry [6] and we are studying the statistics of the spike train output (see, e.g., the software ENAS http://enas.gforge.inria.fr/v3/). Real cell recordings are also analysed in collaboration with Institut de la Vision (Paris). At the level of V1-MT cortical areas, we are investigating the temporal dynamics of motion integration for a wide range of visual stimuli [38], [41], [23], [39], [42] [75], [62][5]. This work is done in collaboration with Institut de Neuroscience 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 [76], coding/decoding approaches [20], [21], [37] [69], [68] or classification [14] [66].

Selected publications on this topic.

PARIETAL Project-Team

3. Scientific Foundations

3.1. Human neuroimaging data and its use

Human neuroimaging consists in acquiring non-invasively image data from normal and diseased human populations. Magnetic Resonance Imaging (MRI) can be used to acquire information on brain structure and function at high spatial resolution.

- T1-weighted MRI is used to obtain a segmentation of the brain into different different tissues, such as gray matter, white matter, deep nuclei, cerebro-spinal fluid, at the millimeter or sub-millimeter resolution. This can then be used to derive geometric and anatomical information on the brain, e.g. cortical thickness.
- Diffusion-weighted MRI measures the local diffusion of water molecules in the brain at the resolution of 2mm, in a set of directions (30 to 60 typically). Local anisotropy, observed in white matter, yields a geometric model of fiber tracts along which water diffusion occurs, and thus provides essential information of the connectivity structure of the brain.
- Functional MRI measures the blood-oxygen-level-dependent (BOLD) contrast that reflects neural activity in the brain, at a spatial resolution of 2 to 3mm, and a temporal resolution of 2-3s. This yields a spatially resolved image of brain functional networks that can be modulated either by specific cognitive tasks or appear as networks of correlated activity.
- Electro- and Magneto-encephalography (MEEG) are two additional modalities that complement functional MRI, as they directly measure the electric and magnetic signals elicited by neural activity, at the millisecond scale. These modalities rely on surface measurements and do not localize brain activity very accurately in the spatial domain.

3.2. High-field MRI

High field MRI as performed at Neurospin (7T on humans, 11.7T in 2013, 17.6T on rats) brings an improvement over traditional MRI acquisitions at 1.5T or 3T, related to to a higher signal-to-noise ratio in the data. Depending on the data and applicative context, this gain in SNR can be traded against spatial resolution improvements, thus helping in getting more detailed views of brain structure and function. This comes at the risk of higher susceptibility distortions of the MRI scans and signal inhomogeneities, that need to be corrected for. Improvements at the acquisition level may come from the use of new coils (such as the new 32 channels coil on the 7T at Neurospin).

3.3. Technical challenges for the analysis of neuroimaging data

The first limitation of Neuroimaging-based brain analysis is the limited Signal-to-Noise Ratio of the data. A particularly striking case if functional MRI, where only a fraction of the data is actually understood, and from which it is impossible to observe by eye the effect of neural activation on the raw data. Moreover, far from traditional i.i.d. Gaussian models, the noise in MRI typically exhibits correlations and long-distance correlation properties (e.g. motion-related signal) and has potentially large amplitude, which can make it hard to distinguish from true signal on a purely statistical basis. A related difficulty is the *lack of salient structure* in the data: it is hard to infer meaningful patterns (either through segmentation or factorization procedures) based on the data only. A typical case is the inference of brain networks from resting-state functional connectivity data.

Regarding statistical methodology, neuroimaging problems also suffer from the relative paucity of the data, i.e. the relatively small number of images available to learn brain features or models, e.g. with respect to the size of the images or the number of potential structures of interest. This leads to several kinds of difficulties, known either as multiple comparison problems or curse of dimensionality. One possibility to overcome this challenge is to increase the amount of data by using images from multiple acquisition centers, at the risk of introducing scanner-related variability, thus challenging the homogeneity of the data. This becomes an important concern with the advent of cross-modal neuroimaging-genetics studies.

SHACRA Project-Team

3. Scientific Foundations

3.1. Biomechanical Modeling

3.1.1. 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 such an approach [24], [27] 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 rely on linearized constitutive laws.

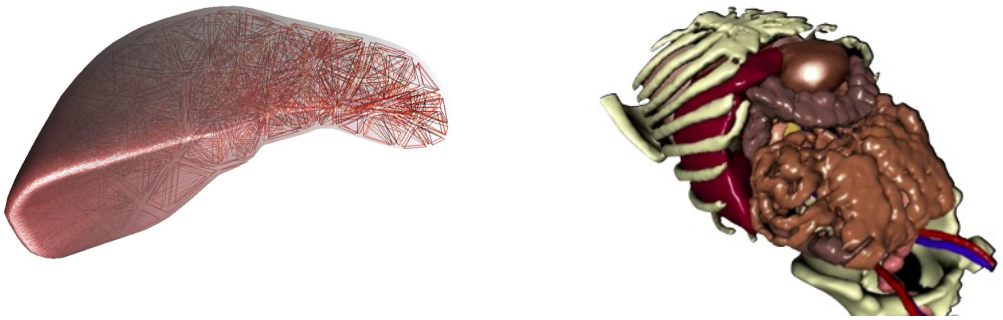


Figure 1. Biomechanical models of organs, based on the Finite Element Method and elasticity theory. Left: a model of the liver based on tetrahedral elements and small strain elasticity. Right: several organ models from a patient dataset combined to create a realistic abdominal anatomy.

An important part of our research is dedicated to the development of new, more accurate models that remain 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. Yet, patient-specific planning or per-operative guidance also requires the models to be parametrized with patient-specific biomechanical data. Very little work has been done in this area, in particular when tissue properties need to be measured in vivo non-invasively. New imaging techniques, such as Ultrasound Elastography or Magnetic Resonance Elastography, could be used to this end [23]. We are currently studying the impact of parametrized patient-specific models of the liver in the context of the PASSPORT european project. This will be used to provide information about the deformation, tissue stiffness and tumor location, for various liver pathologies.

3.1.2. Biomechanical modeling of hollow structures

A large number of anatomical structures in the human body are vascularized (brain, liver, heart, kidneys, ...) and recent interventions (such as interventional radiology) rely on the vascular network as a therapeutical pathway. It is therefore essential to model the shape and deformable behavior of blood vessels. This will be

done at two levels. Global deformation of a vascular network: we have demonstrated previously [9] that we could recover the shape of thousands of vessels from medical images by extracting the centerline of each vessel (see Figure 2). The resulting vascular skeleton can be modeled as a deformable (tree) structure which can capture the global aspects of the deformation. More local deformations can then be described by considering now the actual local shape of the vessel. Other structures such as aneurysms, the colon or stomach can also benefit from being modeled as deformable structures. For this we will rely on shell or thin plate theory. We have recently obtained very encouraging results in the context of the Ph.D. thesis of Olivier Comas [26]. Such local and global models of hollow structures will be particularly relevant for planning coil deployment or stent placement, but also in the context of a new laparoscopic technique called NOTES which uses a combination of a flexible endoscope and flexible instruments. Obtaining patient-specific models of vascular structures and associated pathologies remains a challenge from an image processing stand point, and this challenge is even greater once we require these models to be adapted to complex computational strategies. To this extend we will pursue our collaboration with the MAGRIT team at Inria (through a PhD thesis starting in January 2010) and the Massachusetts General Hospital in Boston.

3.1.3. Blood Flow Simulation

Beyond biomechanical modeling of soft tissues, an essential component of a simulation is the modeling of the functional interactions occurring between the different elements of the anatomy. This involves for instance modeling physiological flows (blood flow, air flow within the lungs...). We particularly plan to study the problem of fluid flow in the context of vascular interventions, such as the simulation of three-dimensional turbulent flow around aneurysms to better model coil embolization procedures. Blood flow dynamics is starting to play an increasingly important role in the assessment of vascular pathologies, as well as in the evaluation of pre- and post-operative status. While angiography has been an integral part of interventional radiology procedures for years, it is only recently that detailed analysis of blood flow patterns has been studied as a mean to assess complex procedures, such as coil deployment. A few studies have focused on aneurysm-related hemodynamics before and after endovascular coil embolization. Groden et al. [31] constructed a simple geometrical model to approximate an actual aneurysm, and evaluated the impact of different levels of coil packing on the flow and wall pressure by solving Navier-Stokes equations, while Kakalis et al. [33] relied on patient-specific data to get more realistic flow patterns, and modeled the coiled aneurysm as a porous medium. As these studies aimed at accurate Computational Fluid Dynamics simulation, they rely on commercial software, and the computation times (dozens of hours in general) are incompatible with interactive simulation or even clinical practice. Generally speaking, accuracy and efficiency are two significant pursuits in numerical calculation, but unfortunately very often contradictory.

With the Ph.D. thesis of Yiyi Wei, we have recently started the development of a new technique for accurately computing, in near real-time, the flow of blood within an aneurysm, as well as the interaction between blood and coils. In this approach we rely on the Discrete Exterior Calculus method to obtain an ideal trade-off between accuracy and computational efficiency. Although still at an early stage, these results show that our approach can accurately capture the main characteristics of the complex blood flow patterns in and around an aneurism. The model also takes into account the influence of the coil on the blood flow within the aneurysm. The main difference between our approach and many other work done by internationally renowned teams (such as REO team at Inria or the Computer Vision Laboratory at ETH) comes from the importance we place in the computational efficiency of the method. To some extent our approach is similar to what has been done to obtain real-time finite element methods. We are essentially trying to capture the key characteristics of the behavior for a particular application. This is well illustrated by the work we started on flow modeling, which received an award in September 2009 at the selective conference on Medical Image Computing and Computer Assisted Interventions [10]. We will pursue this direction to accurately model the local flow in a closed domain (blood vessel, aneurysm ventricle, ...) and combine it with some of our previous work describing laminar flow across a large number of vessels [38] in order to define boundary conditions for the three-dimensional model.

3.2. Biomechanical Systems

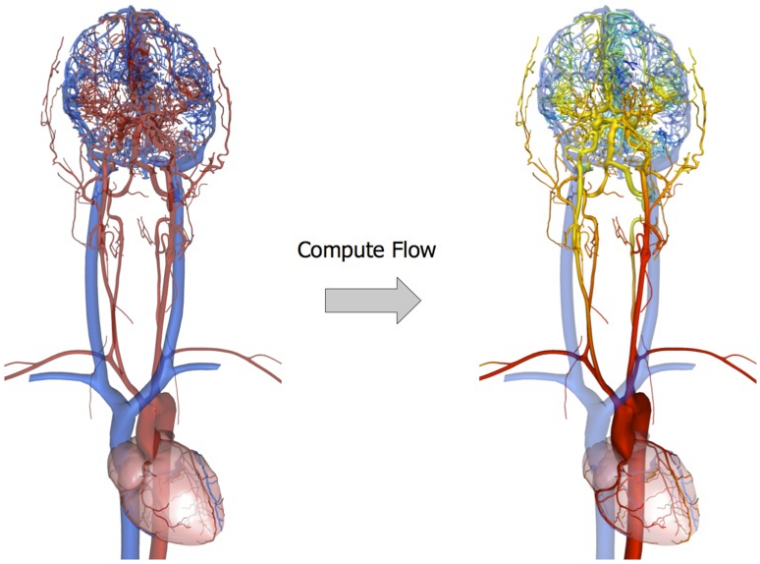


Figure 2. Blood flow and pressure distribution in the cerebrovascular system. The arterial vascular network is composed of more than 3,000 vessels, yet the computation is performed in real-time.

3.2.1. Constraint models and boundary conditions

To accurately model soft tissue deformations, the approach must account for the intrinsic behavior of the target organ, but also for its biomechanical interactions with surrounding tissues or with medical devices. While the biomechanical behavior of important organs (such as the brain or liver) has been well studied, few work exists regarding the mechanical interactions between the anatomical structures. For tissue-tool interactions, most approaches rely on a simple contact models, and rarely account for friction. While this simplification can produce plausible results in the case of an interaction between the end effector of a laparoscopic instrument and the surface of an organ, it is generally an incorrect approximation. As we move towards simulations for planning or rehearsal, accurately modeling contacts will take an increasingly important place. We have recently shown in [28] and [29] that we could compute, in real-time, complex interactions between a coil and an aneurysm, or between a flexible needle and soft-tissues. In laparoscopic surgery, the main challenge lies in the modeling of interactions between anatomical structures rather than between the instruments and the surface of an organ. During the different steps of a procedure organs slides against each other, while respiratory, cardiac and patient motion also generate contacts. Modeling these multiple interactions becomes even more complex when different biomechanical models are used to characterize the various soft tissues of the anatomy. Consequently, our objective is to accurately model resting contacts with friction, in a heterogeneous environment (spring-mass models, finite element models, particle systems, rigid objects, etc.). When different time integration strategies are used, a challenge lies in the computation of contact forces in a way that integrity and stability of the overall simulation are maintained. Our objective is to work on the definition of these various boundary conditions and on new resolution methods for such heterogeneous simulations. In particular we will investigate a simulation process in which each model continues to benefit from its own optimizations while taking into account the mechanical couplings due to interactions between objects.

3.2.2. Vascularized anatomy

From a clinical standpoint, several procedures involve vascularized anatomical structures such as the liver, the kidneys, or the brain. When a therapy needs to be applied on such structures, it is currently possible to perform

a procedure surgically or to use an endovascular approach. This requires to characterize and model the behavior of vessels (arteries and veins) as well as the behavior of soft tissue (in particular the parenchyma). Another challenge of this research will be to model the interactions between the vascular network and the parenchyma where it is embedded. These interactions are key for both laparoscopic surgery and interventional radiology as they allow to describe the motion of the vessels in a vascularized organ during the procedure. This motion is either induced by the surgical manipulation of the parenchymal tissue during surgery or by respiratory, cardiac or patient motion during interventional radiology procedures. From a biomechanical standpoint, capillaries are responsible for the viscoelastic behavior of the vascularized structures, while larger vessels have a direct impact on the overall behavior of the anatomy. In the liver for instance, the apparent stiffness of the organ changes depending on the presence or absence of large vessels. Also, the relatively isotropic nature of the parenchyma is modified around blood vessels. We propose to model the coupling that exists between these two different anatomical structures to account for their respective influence. For this we will initially rely on the work done during the Ph.D. thesis of Christophe Guebert (see ([32] for instance) and we will also investigate coupling strategies based on degrees of freedom reduction to reduce the complexity of the problem (and therefore also computation times). Part of this work is already underway in the context of the PASSPORT european project with IRCAD and soft tissue measurements will be performed in collaboration with the biomechanics laboratory at Strasbourg University.

3.2.3. Parallel Computation

Although the past decade has seen a significant increase in complexity and performance of the algorithms used in medical simulation, major improvements are still required to enable patient-specific simulation and planning. Using parallel architectures to push the complexity of simulated environments further is clearly an approach to consider. However, interactive simulations introduce new constraints and evaluation criteria, such as latencies, multiple update frequencies and dynamic adaptation of precision levels, which require further investigation. New parallel architectures, such as multi-cores CPUs, are now ubiquitous as the performances achieved by sequential units (single core CPUs) stopped to regularly improve. At the same time, graphical processors (GPU) offer a massive computing power that is now accessible to non-graphical tasks thanks to new general-purposes API such as CUDA and OpenCL. GPUs are internally parallel processors, exploiting hundreds of computing units. These architectures can be exploited for more ambitious simulations, as we already have demonstrated in a first step by adding support for CUDA within the SOFA framework. Several preliminary results of GPU-based simulations have been obtained, permitting to reach speedup factors (compared to a single core GPU) ranging from 16x to 55x. Such improvements permit to consider simulations with finer details, or new algorithms modeling biomechanical behaviors more precisely. However, while the fast evolution of parallel architectures is useful to increase the realism of simulations, their varieties (multi-core CPUs, GPUs, clusters, grids) make the design of parallel algorithm challenging. An important effort needs to be made is to minimize the dependency between simulation algorithms and hardware architectures, allowing the reuse of parallelization efforts on all architecture, as well as simultaneously exploiting all available computing resources present in current and future computers. The largest gains could be achieved by combining parallelism and adaptive algorithms. The design and implementation of such a system is a challenging problem, as it is no longer possible to rely on pre-computed repartition of datas and computations. Thus, further research is required in highly adaptive parallel scheduling algorithms, and highly efficient implementation able to handle both large changes in computational loads due to user interactions and multi-level algorithms, and new massively parallel architectures such as GPUs. A direction that we are also investigating is to combine multi-level representations and locally adaptive meshes. Multi-level algorithms are useful not only to speedup computations, but also to describe different characteristics of the deformation at each level. Combined with local change of details of the mesh (possibly using hierarchical structures), the simulation can reach a high level of scalability.

VISAGES Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

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 tumour imaging for instance).
- In the field of image analysis and statistical modelling 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 labelling 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.

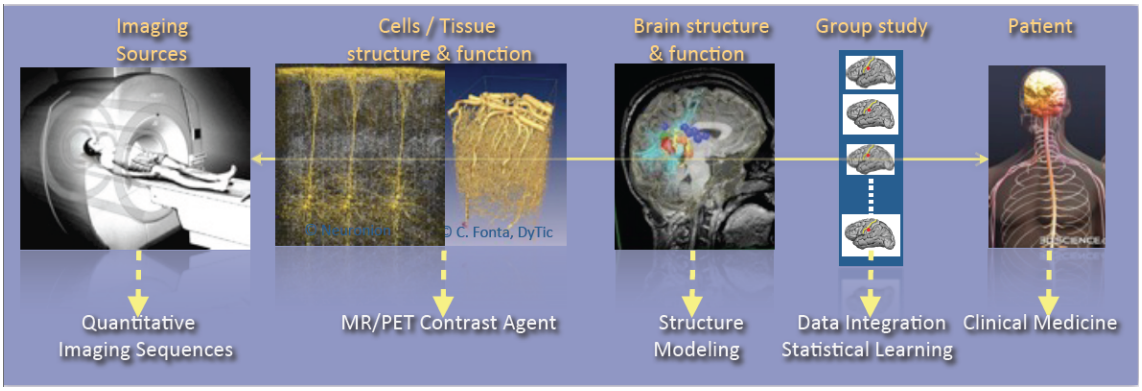


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales : from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions,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

CLIME Project-Team

3. Scientific Foundations

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 developed by CEREA: 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 initiatives.

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 value. 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 MétéoFrance, 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).

FLUMINANCE Project-Team

3. Scientific Foundations

3.1. Fluid flow analysis and modeling

Turbulent fluid flows involved in environmental or industrial applications are complex. In fluid mechanics laboratories, canonical turbulent shear flows have been studied for many years and a relatively clear picture of their underlying structure exists. However, the direct applicability of these efforts to real relevant flows, which often occur in complex geometries and in the presence of multiple non canonical influences, like cross-shear, span wise non-uniform and thermal stratification, is still unknown. In addition, the turbulence can be characterized by Reynolds number ranging between 10^3 and 10^4, corresponding to transitional regime for which the use of classical turbulence models is limited.

In this context, we have performed research studies on turbulent shear flows of low velocities by tackling crucial topics of measurements, analysis and modeling of environmental and industrial flows in presence of non-canonical influences. This concerns more precisely the study of the interaction between a mixing layer and circular cylinder wake flow, the study of wake flow with span wise non uniformity, the study of mixing layer under the influence of thermal stratification and the study of mixing layer forced between non-uniform flows. The analysis of these flows has required the design of adequate dynamical models, using proper orthogonal decomposition and Galerkin projection. Understanding issues such as the mechanisms of heat and mass transfer involved in these shear flows provides meaningful information for the control of relevant engineering flows and the design of new technologies. To investigate more thoroughly these complex flows numerical and experimental tools have been designed. An immersed boundary method was proposed to mimic complex geometries into Direct Numerical simulation (DNS) and Large Eddy Simulations (LES) codes. A novel anemometer has been designed and implemented for the simultaneous measurement of velocity and temperature in air flows with a single hot-wire probe.

Mixing layer wake interaction
We have investigated the vortex shedding of a circular cylinder immersed in a plane turbulent mixing layer. For a centre span Reynolds number of 7500, the wake flow splits into three regions: a high-velocity wake, a low-velocity wake and a region of interaction in the middle span of the body. A strong unsteady secondary flow is observed, and explained with span wise base pressure gradients. Unexpected features are found for formation length and the base pressure along the span of the cylinder. In the high-velocity side, where the local Reynolds number is the highest, the formation length is longest. Based on the formation length measurements it was shown that as a function of the centre span Reynolds number, the wake flows behaves as circular cylinder in uniform flow. Three cells with a constant frequency with adjacent dislocations are observed. For each cell, a shedding mode was suggested. The relation of the secondary flow to the frequencies was examined. All the observations were analyzed by analogical reasoning with other flows. This pointed out the action of the secondary flow in the high-velocity side regarded as a wake interference mechanism.

Low order complex flow modeling
We have proposed improvements to the construction of low order dynamical systems (LODS) for incompressible turbulent external flows. The reduced model is obtained by means of a Proper Orthogonal Decomposition (POD) basis extracted through a truncated singular value decomposition of the flow auto-correlation matrix built from noisy PIV experimental velocity measurements. The POD modes are then used to formulate a reduced dynamical system that contains the main features of the flow. This low order dynamical system (LODS) is obtained through a Galerkin projection of the Navier-Stokes Equations on the POD basis. Usually, the resulting system of ordinary differential equations presents stability problems due to modes truncation and numerical uncertainties, especially when working on experimental data. The technique we proposed relies on an optimal control approach to estimate the dynamical system coefficients and its initial condition. This allows us to recover a reliable and stable spatio-temporal reconstruction of the large scales of the flow. The technique

has been assessed on the near wake behind a cylinder observed through very noisy PIV measurement. It has been also evaluated for configurations involving a rotating cylinder.

Studies on complex 3D dynamical behavior resulting from the interaction between a plane mixing layer and the wake of a cylinder have been also investigated using POD representation, applied to data from two synchronized 2D PIV systems (Dual-plane PIV). This approach allowed us to construct a 3D-POD representation. An analysis of the correlations shows different length scales in the regions dominated by wake like structures and shear layer type structures [2]. In order to characterize the particular organization in the plane of symmetry, a Galerkin projection from a slice POD has been performed. This led to a low-dimensional dynamical system that allowed the analysis of the relationship between the dominant frequencies. This study led to a reconstruction of the dominant periodic motion suspected from previous studies [41]. This work allowed us to make a link between the three-dimensional organization and the secondary unsteady motion from the low velocity side to the high velocity side of the mixing layer, appearing in this highly 3D flow configuration.

Direct and Large Eddy simulations of complex flows

We have proposed a direct forcing method better suited to the use of compact finite difference schemes in Direct Numerical Simulation. The new forcing creates inside the body an artificial flow preserving the no-slip condition at the surface but reducing the step-like change of the velocity derivatives across the immersed boundary. This modification led to improve results both qualitatively and quantitatively for conventional and complex flow geometries [50].

Three-dimensional direct numerical simulations have been performed for vortex shedding behind cylinders. We focused in particular on cases for which the body diameter and the incoming flow involved span wise linear non-uniformity. Four configurations were considered: the shear flow, the tapered cylinder and their combinations, which gave rise namely to the adverse and aiding cases. In contrast with the observations of other investigators, these computations highlighted distinct vortical features between the shear case and the tapered case. In addition, it was observed that the shear case and the adverse case (respectively the tapered and aiding case), yielded similarities in flow topology. This phenomenon was explained by the span wise variations of the ratio of mean velocity and the cylinder diameter which seemed to govern these flows. Indeed, it was observed that large span wise variations of U/D seemed to enhance three-dimensionality, through the appearance of vortex-adhesions and dislocations. Span wise cellular pattern of vortex shedding were identified. Their modifications in cell size, junction position and number were correlated with the variation of U/D. In the Lee side of the obstacle a wavy secondary motion was identified. Induced secondary flow due to the bending of Karman vortices in the vicinity of vortex-adhesion and dislocations was suggested to explain this result [49].

LES and experimental wake flow database

We contributed to the study of flow over a circular cylinder at Reynolds number $Re = 3900$. Although this classical flow is widely documented in the literature, especially for this precise Reynolds number, which leads to a sub critical flow regime, there is no consensus about the turbulence statistics immediately just behind the obstacle. This flow has been studied both numerically with Large Eddy Simulation and experimentally with Hot-Wire Anemometry and Particle Image Velocimetry. The numerical simulation has been performed using high-order schemes and the specific Immersed Boundary Method previously mentioned. We focused on turbulence statistics and power spectra in the near wake up to 10 diameters. Statistical estimation is shown to need large integration times increasing the computational cost and leading to an uncertainty of about 10% for most flow characteristics considered in this study. The present numerical and experimental results are found to be in good agreement with previous Large Eddy Simulation data. Our study has exhibited significant differences compared with the experimental data found in the literature. The obtained results attenuate previous numerical-experimental controversy for this type of flows [11].

Simultaneous velocity temperature measurements in turbulent flows

We have worked on the design of a novel anemometer for the simultaneous measurement of velocity and temperature in airflows with a single hot wire probe. The principle of periodically varying the overheat ratio of the wire has been selected and applied through a tunable electronic chain. Specific methods were developed for the calibration procedure and the signal processing. The accuracy of the measurements was assessed by means of Monte-Carlo simulations. Accurate results were provided for two types of turbulent non-isothermal flows, a coaxial heated jet and a low speed thermal mixing. The particular interest of the synchronization of the two measurements has been emphasized during the PhD thesis of T. Ndoye.

A new dynamic calibration technique has been developed for hot-wire probes. The technique permits, in a short time range, the combined calibration of velocity, temperature and direction calibration of single and multiple hot-wire probes. The calibration and measurements uncertainties were modeled, simulated and controlled, in order to reduce their estimated values.

3.2. Fluid motion analysis

Flow visualization has been a powerful tool to depict or to understand flow feature properties. Efforts to develop high-quality flow visualization techniques date back over a century. The analysis of the recorded images consisted firstly to a qualitative interpretation of the streak lines leading to an overall global insight into the flow properties but lacking quantitative details on important parameters such as velocity fields or turbulence intensities. Point measurement tools such as hot wire probes or Laser Doppler Velocimetry have typically provided these details. As these probes give information only at the point where they are placed, simultaneous evaluations at different points require to dispose a very large number of probes and the evaluation of unsteady field (most of the flows are unsteady) is almost unachievable with them.

In an effort to avoid the limitations of these probes, the Particle Image Velocimetry (PIV), a non-intrusive diagnostic technique, has been developed in the last two decades [40]. The PIV technique enables obtaining velocity fields by seeding the flow with particles (e.g. dye, smoke, particles) and observing the motion of these tracers. In computer vision, the estimation of the projection of the apparent motion of a 3D scene onto the image plane, refereed 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 [45]. Unlike to dense optical flow estimators, the former approach provides techniques that supply only sparse velocity fields. These methods have demonstrated to be robust and to provide accurate measurements for flows seeded with particles. These restrictions and their inherent discrete local nature limit too much their use and prevent any evolutions of these techniques towards the devising of methods supplying physically consistent results and small scale velocity measurements. It does not authorize also the use of scalar images exploited in numerous situations to visualize flows (image showing the diffusion of a scalar such as dye, pollutant, light index refraction, flurocein,...). At the opposite, variational techniques enable in a well-established mathematical framework to estimate spatially continuous velocity fields, which should allow more properly to go towards the measurement of smaller motion scales. As these methods are defined through PDE's systems they allow quite naturally including as constraints the kinematical and dynamical laws governing the observed fluid flows. Besides, within this framework it is also much easier to define characteristic features estimation 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. This route has demonstrated to be much more robust to scalar image. Several studies in this vein have strengthened our skills in this domain. All the following approaches have been either formulated within a statistical Markov Random Fields modeling or either within a variational framework. For a thorough description of these approaches see [7].

ICE data model and div-curl regularization This fluid motion estimator is constructed on a data model derived from the Integration of the Continuity Equation (ICE data model) [5] and includes a second order regularization scheme enabling to preserve blobs of divergence and curl. Intensive evaluations of this estimator on flow prototypes mastered in laboratory have shown that this estimator led to the same order of accuracy as the best PIV techniques but for an increase information density. This ability to get dense flow fields allowed us estimating proper vorticity or divergence maps without resorting to additional post-processing interpolation schemes.

Schlieren Image velocimetry We have addressed the problem of estimating the motion of fluid flows visualized with the Schlieren technique. Such an experimental visualization system is well known in fluid mechanics and it enables the visualization of unseeded flows. This technique authorizes the capture of phenomena that are impossible to visualize with particle seeding such as natural convection, phonation flow, breath flow and allows the setting of large scale experiments. Since the resulting images exhibit very low intensity contrasts, classical motion estimation methods based on the brightness constancy assumption (correlation-based approaches, optical flow methods) are completely inefficient. The global energy function we have defined for Schlieren images is composed of i) a specific data model accounting for the fact that the observed luminance is related to the gradient of the fluid density, and ii) a specific constrained div-curl regularization term. To date there exists no motion estimator allowing estimating accurately dense velocity fields on Schlieren images.

Low order fluid motion estimator This low-dimensional fluid motion estimator [6] is based on the Helmholtz decomposition, which consists in representing the velocity field as the sum of a divergence-free component and a curl-free one. In order to provide a low-dimensional solution, both components have been approximated using a discretization of the vorticity (curl of the velocity vector) and divergence maps through regularized Dirac measures [44]. The resulting so-called irrotational (resp. solenoidal) field is then represented by a linear combination of basis functions obtained by a convolution product of the Green kernel gradient and the vorticity map (resp. the divergence map). The coefficient values and the basis function parameters are obtained by minimizing a function formed by an integrated version of the mass conservation principle of fluid mechanics.

Potential functions estimation and finite mimetic differences We have studied a direct estimation approach of the flow potential functions (respectively the *stream* function and the *velocity* potential) from two consecutive images. The estimation has been defined on the basis of a high order regularization scheme and has been implemented through mimetic difference methods[12]. With these approaches the discretization preserves basic relationships of continuous vector analysis. Compared to previous discretization scheme based on auxiliary div-curl variables, the considered technique appeared to be numerically much more stable and led to an improve accuracy.

2D and 3D atmospheric motion layer estimation In this study, we have explored the problem of estimating mesoscales dynamics of atmospheric layers from satellite image sequences. Due to the intrinsic sparse 3-dimensional nature of clouds and to large occluded zones caused by the successive overlapping of cloud layers, the estimation of accurate layered dense motion fields is an intricate issue. Relying on a physically sound vertical decomposition of the atmosphere into layers, we have proposed two dense motion estimators for the extraction of multi-layer horizontal (2D) and 3D wind fields. These estimators are expressed as the minimization of a global function that includes a data-driven term and a spatio-temporal smoothness term. A robust data term relying on shallow-water mass conservation model has been proposed to fit sparse observations related to each layer. In the 3D case, the layers are interconnected through a term modeling mass exchanges at the layers surfaces frontiers [9].

A novel spatio-temporal regularizer derived from the shallow-water momentum conservation model has been considered to enforce temporal consistency of the solution along time. These constraints are combined with a robust second-order regularizer preserving divergent and vorticity structures of the flow. Besides, a two-level motion estimation scheme has been settled to overcome the limitations of the multiresolution incremental estimation scheme when capturing the dynamics of fine mesoscale structures. This alternative approach relies on the combination of correlation and optical-flow observations. An exhaustive evaluation of the novel method has been first performed on a scalar image sequence generated by Direct Numerical Simulation of a turbulent bi-dimensional flow. Based on qualitative experimental comparisons, the method has also been assessed on a Meteosat infrared image sequence.

3.3. Data assimilation and Tracking of characteristic fluid features

Classical motion estimation techniques usually proceed on pairs of two successive images, and do not enforce temporal consistency. This often induces an estimation drift which is essentially due to the fact that motion estimation is formulated as a local process in time. No adequate physical dynamics law, or conservation law,

related to the observed flow, is taken into account over long time intervals by the usual motion estimators. The estimation of an unknown state variable trajectory on the basis of specified dynamical laws and some incomplete and noisy measurements of the variable of interest can be either conducted through optimal control techniques or through stochastic filtering approach. These two frameworks have their own advantages and deficiencies. We rely indifferently on both approaches.

Stochastic filtering for fluid motion tracking We have proposed a recursive Bayesian filter for tracking velocity fields of fluid flows. The filter combines an Îto diffusion process associated to 2D vorticity-velocity formulation of Navier-Stokes equation and discrete image error reconstruction measurements. In contrast to usual filters, designed for visual tracking problem, our filter combines a continuous law for the description of the vorticity evolution with discrete image measurements. We resort to a Monte-Carlo approximation based on particle filtering. The designed tracker provides a robust and consistent estimation of instantaneous motion fields along the whole image sequence. In order to handle a state space of reasonable dimension for the stochastic filtering problem, the motion field is represented as a combination of adapted basis functions. The basis functions are derived from a mollification of Biot-Savart integral and a discretization of the vorticity and divergence maps of the fluid vector field. The output of such a tracking is a set of motion fields along the whole time range of the image sequence. As the time discretization is much finer than the frame rate, the method provides consistent motion interpolation between consecutive frames. In order to reduce further the dimensionality of the associated state space when we are facing a large number of motion basis functions, we have explored a new dimensional reduction approach based on dynamical systems theory. The study of the stable and unstable directions of the continuous dynamics enables to construct an adaptive dimension reduction procedure. It consists in sampling only in the unstable directions, while the stable ones are treated deterministically [6].

When the likelihood of the measurement can be modeled as Gaussian law, we have also investigated the use of so-called ensemble Kalman filtering for fluid tracking problems. This kind of filters introduced for the analysis of geophysical fluids is based on the Kalman filter update equation. Nevertheless, unlike traditional Kalman filtering setting, the covariances of the estimation errors, required to compute the Kalman gain, rely on an ensemble of forecasts. Such a process gives rise to a Monte Carlo approximation for a family of stochastic non linear filters enabling to handle state spaces of large dimension. We have recently proposed an extension of this technique that combines sequential importance sampling and the propagation law of ensemble kalman filter. This technique leads to ensemble Kalman filter with an improve efficiency. This appears to be a generalization of the optimal importance sampling strategy we proposed in the context of partial conditional Gaussian trackers [1].

Variational assimilation technique

We investigated the use of variational framework for the tracking from image sequence of features belonging to high dimensional spaces. This framework relies on optimal control principles as developed in environmental sciences to analyze geophysical flows [46], [47]. Within the PhD of Nicolas Papadakis [10], we have first devised a data assimilation technique for the tracking of closed curves and their associated motion fields. The proposed approach enables a continuous tracking along an image sequence of both a deformable curve and its associated velocity field. Such an approach has been formalized through the minimization of a global spatio-temporal continuous cost functional, with respect to a set of variables representing the curve and its related motion field. The resulting minimization sequence consists in a forward integration of an evolution law followed by a backward integration of an adjoint evolution model. The latter pde includes a term related to the discrepancy between the state variables evolution law and discrete noisy measurements of the system. The closed curves are represented through implicit surface modeling [48], whereas the motion is described either by a vector field or through vorticity and divergence maps according to the type of targeted application. The efficiency of the approach has been demonstrated on two types of image sequences showing deformable objects and fluid motions.

More recently assimilation technique for the direct estimation of atmospheric wind field from pressure images have been proposed [4]. These techniques rely on a brightness variation model of the intensity function. They do not include anymore motion measurements provided by external motion estimators. The resulting estimator

allows us to recover accurate fluid motion fields and enables tracking dense vorticity maps along an image sequence.

3.4. Visual servoing

Nowadays, visual servoing is a widely used technique is robot control. It consists in using data provided by a vision sensor for controlling the motions of a robot [43]. Various sensors can be considered such as perspective cameras, omnidirectional cameras, 2D ultrasound probes or even virtual cameras. In fact, this technique is historically embedded in the larger domain of sensor-based control [51] so that other sensors than vision sensors can be properly used. On the other hand, this approach was first dedicated to robot arms control. Today, much more complex system can be considered like humanoid robots, cars, submarines, airships, helicopters, aircrafts. Therefore, visual servoing is now seen as a powerful approach to control the state of dynamic systems.

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 the image. A control law is then designed so that these visual features reach a desired value $\mathbf{s}^*$ related to the desired state of the system. The control principle is thus to regulate 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 camera instantaneous velocity $\mathbf{v}$

$$\dot{\mathbf{s}} = \mathbf{L_s}\,\mathbf{v} + \frac{\partial \mathbf{s}}{\partial t} \tag{10}$$

where the term $\frac{\partial \mathbf{s}}{\partial t}$ describes the non-stationary behavior of $\mathbf{s}$. Typically, if we try to ensure an exponential decoupled decrease of the error signal and if we consider the camera velocity as the input of the robot controller, the control law writes as follow

$$\mathbf{v} = -\lambda \widehat{\mathbf{L_s}}^{+}\mathbf{e} - \widehat{\mathbf{L_s}}^{+}\frac{\widehat{\partial \mathbf{e}}}{\partial t} \tag{11}$$

with λ a proportional gain that has to be tuned to minimize the time-to-convergence, $\widehat{\mathbf{L_s}}^{+}$ the pseudo-inverse of a model or an approximation of $\mathbf{L_s}$ and $\frac{\widehat{\partial \mathbf{e}}}{\partial t}$ an estimation of $\frac{\partial \mathbf{e}}{\partial t}$.

The behavior of the closed-loop system is then obtained, from (2), by expressing the time variation of the error $\mathbf{e}$

$$\dot{\mathbf{e}} = -\lambda \mathbf{L_s}\widehat{\mathbf{L_s}}^{+}\mathbf{e} - \mathbf{L_s}\widehat{\mathbf{L_e}}^{+}\frac{\widehat{\partial \mathbf{e}}}{\partial t} + \frac{\partial \mathbf{e}}{\partial t}. \tag{12}$$

As can be seen, visual servoing explicitly relies on the choice of the visual features $\mathbf{s}$ and then on the related interaction matrix; that is the key point of this approach. Indeed, this choice must be performed very carefully. Especially, an isomorphism between the camera pose and the visual features is required to ensure that the convergence of the control law will lead to the desired state of the system. An optimal choice would result in finding visual features leading to a diagonal and constant interaction matrix and, consequently, to a linear decoupled system for which the control problem is well known. Thereafter, the isomorphism as well as the global stability would be guaranteed. In addition, since the interaction matrix would present no more nonlinearities, a suitable robot trajectory would be ensured.

However, finding such visual features is a very complex problem and it is still an open issue. Basically, this problem consists in building the visual features $\mathbf{s}$ from the nonlinear visual measurements $\mathbf{m}$ so that the interaction matrix related to $\mathbf{s}$ becomes diagonal and constant or, at least, as simple as possible.

On the other hand, a robust extraction, matching (between the initial and desired measurements) and real-time spatio-temporal tracking (between successive measurements) have to be ensure but have proved to be a complex task, as testified by the abundant literature on the subject. Nevertheless, this image process is, to date, a necessary step and often considered as one of the bottlenecks of the expansion of visual servoing. That is why more and more non geometric visual measurements are proposed [3].

3.5. Sparse Representations and Bayesian model selection

Sparse representation methods aim at finding representations of a signal with a small number of components taken from an over-complete dictionary of elementary functions or vectors. Sparse representation are of interest in a number of applications in Physics and signal processing. In particular, they provide a simple characterization of certain families of signals encountered in practice. For example, smooth signals can be shown to have a sparse representation in over-complete Fourier or wavelet dictionaries. More recently, it has been emphasized in [42] that the solutions of certain differential equations (*e.g.,* diffusion or transport equation) have a sparse representation in dictionaries made up of curvelets.

Finding the sparse representation of a signal typically requires to solve an under-determined system of equations under the constraint that the solution is composed of the minimum number of non-zero elements. Unfortunately, this problem is known to be NP-hard and sub-optimal procedures have to be devised to find practical solutions. Among the various algorithms that find approximate solutions, let us mention for example the matching pursuit, orthogonal matching pursuit or basis pursuit algorithms.

Choosing appropriate models and fixing hyper-parameters is a tricky and often hidden process in optic-flow estimation. Most of the motion estimators proposed so far have generally to rely on successive trials and a empirical strategy for fixing the hyper parameters values and choosing the adequate model. Besides of its computational inefficiency, this strategy may produce catastrophic estimate without any relevant feedback for the end-user, especially when motions are difficult to apprehend as for instance for complex deformations or non-conventional imagery. Imposing hard values to these parameters may also yield poor results when the lighting conditions or the underlying motions differ from those the system has been calibrated with. At the extreme, the estimate may be either too smooth or at the opposite non-existent strong motion discontinuities.

Bayesian model selection offers an attractive solution to this problem. The Bayesian paradigm implicitly requires the definition of several competing observation and prior *probabilistic* model(s). The observation model relates the motion of the physical system to the spatial and temporal variations of the image intensity. The prior models define the spatio-temporal constraints that the motion have to satisfy. Considering these competing models, the Bayesian theory provides methodologies to select the best models under objective performance criterion (minimum probability of error, minimum mean square error, etc). Moreover, due to the generality of this problem, numerous algorithms and approximations exist in the literature to implement efficient and effective practical solutions: Monte-Carlo integration, mean-field and Laplace approximations, EM algorithm, graphical models, etc.

MAGIQUE-3D Project-Team

3. Scientific Foundations

3.1. Inverse Problems

- **Inverse scattering problems.** The determination of the shape of an obstacle from its effects on known acoustic or electromagnetic waves is an important problem in many technologies such as sonar, radar, geophysical exploration, medical imaging and nondestructive testing. This inverse obstacle problem (IOP) is difficult to solve, especially from a numerical viewpoint, because it is ill-posed and nonlinear [77]. Moreover the precision in the reconstruction of the shape of an obstacle strongly depends on the quality of the given far-field pattern (FFP) measurements: the range of the measurements set and the level of noise in the data. Indeed, the numerical experiments (for example [88], [91], [84], [85]) performed in the resonance region, that is, for a wavelength that is approximately equal to the diameter of the obstacle, tend to indicate that in practice, and at least for simple shapes, a unique and reasonably good solution of the IOP can be often computed using only one incident wave and *full aperture* far-field data (FFP measured only at a limited range of angles), as long as the aperture is larger than π. For smaller apertures the reconstruction of the shape of an obstacle becomes more difficult and nearly impossible for apertures smaller than $\pi/4$.
  This plus the fact that from a mathematical viewpoint the FFP can be determined on the entire sphere $S1$ from its knowledge on a subset of $S1$ because it is an *analytic* function, we propose [74], [75] a solution methodology to extend the range of FFP data when measured in a limited aperture and not on the entire sphere $S1$. It is therefore possible to solve the IOP numerically when only limited aperture measurements are available. The objective of MAGIQUE-3D is to extend this work to 3D problems of acoustic scatterin and to tackle the problem of elasto-acoustic scattering.
- **Depth Imaging in the context of DIP.** The challenge of seismic imaging is to obtain the best 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 most used technique of imaging is RTM (Reverse Time Migration), [76], which 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. The research program DIP has been defined by researchers of MAGIQUE-3D and engineers of TOTAL jointly. It has been created with the aim of gathering researchers of Inria, with different backgrounds and the scientific programm will be coordinated by MAGIQUE-3D. In this context, MAGIQUE-3D will contribute by working on the inverse problem and by continuing to develop 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 focus 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 Schemes in Space and Time.** 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 considerably reduce the space step, which is the distance between two points of the mesh representing the computational domain. Obviously this results in increasing the number of unknowns of the discrete problem. Besides, the time step, whose value fixes the number of required iterations for solving the evolution problem, is linked to the space step through

the 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 (and of multiplications by the stiffness matrix) will be. The method that we proposed in [11] allows for the use of local time-step, adapted to the various sizes of the cells and we recently extended it to deal with p-adaptivity [73]. However, this method can not yet handle dissipation terms, which prevents us for using absorbing boundary conditions or Perfectly Matched Layers (PML). To overcome this difficulty, we will first tackle the problem to used the modified equation technique [78], [90], [82] with dissipation terms, which is still an open problem.

We are also considering an alternative approach to obtain high-order schemes. The main idea is to apply first the time discretization thanks to the modified equation technique and after to consider the space discretization. Our approach involves p-harmonic operators, which can not be discretized by classical finite elements. For the discretization of the biharmonic operator in an homogeneous acoustic medium, both C1 finite elements (such as the Hermite ones) and Discontinuous Galerkin Finite Elements (DGFE) can be used while in a discontinuous medium, or for higher-order operators, DGFE should be preferred [68]. This new method seems to be well-adapted to p-adaptivity. Therefore, we now want to couple it to our local time-stepping method in order to deal with hp-adaptivity both in space and in time. We will then carry out theoretical and numerical comparisons between this technique and the classical modified equation scheme.

Once we have performant hp-adaptive techniques, it will be necessary to obtain error-estimators. Since we consider huge domain and complex topography, the remeshing of the domain at each time-step is impossible. One solution would be to remesh the domain for instance each 100 time steps, but this could also hamper the efficiency of the computation. Another idea is to consider only p adaptivity, since in this case there is no need to remesh the domain.

- **Mixed hybrid finite element methods for the wave equation.** The new mixed-hybrid-like method for the solution of Helmholtz problems at high frequency we have built enjoys the three following important properties: (1) unlike classical mixed and hybrid methods, the method we proposed is not subjected to an inf-sup condition. Therefore, it does not involve numerical instabilities like the ones that have been observed for the DGM method proposed by Farhat and his collaborators [80], [81]. We can thus consider a larger class of discretization spaces both for the primal and the dual variables. Hence we can use unstructured meshes, which is not possible with DGM method (2) the method requires one to solve Helmholtz problems which are set inside the elements of the mesh and are solved in parallel(3) the method requires to solve a system whose unknowns are Lagrange multipliers defined at the interfaces of the elements of the mesh and, unlike a DGM, the system is hermitian and positive definite. Hence we can use existing numerical methods such as the gradient conjugate method. We intend to continue to work on this subject and our objectives can be described following three tasks: (1) Follow the numerical comparison of performances of the new methods with the ones of DGM. We aim at considering high order elements such as R16-4, R32-8, ...; (2)Evaluate the performance of the method in case of unstructured meshes. This analysis is very important from a practical point of view but also because it has been observed that the DGM deteriorates significantly when using unstructured meshes; (3) Extend the method to the 3D case. This is the ultimate objective of this work since we will then be able to consider applications.

  Obviously the study we propose will contain a mathematical analysis of the method we propose. The analysis will be done in the same time and we aim at establishing a priori and a posteriori estimates, the last being very important in order to adopt a solution strategy based on adaptative meshes.

- **Boundary conditions.** The construction of efficient absorbing conditions is very important for solving wave equations, which are generally set in unbounded or very large domains. The efficiency of the conditions depends on the type of waves which are absorbed. Classical conditions absorb propagating waves but recently new conditions have been derived for both propagating and evanescent waves in the case of flat boundaries. MAGIQUE-3D would like to develop new absorbing boundary conditions whose derivation is based on the full factorization of the wave equation using pseudodifferential calculus. By this way, we can take the complete propagation phenomenon into account

which means that the boundary condition takes propagating, grazing and evanescent waves into account, and then the absorption is optimized. Moreover our approach can be applied to arbitrarily-shaped regular surfaces.
We intend to work on the development of interface conditions that can be used to model rough interfaces. One approach, already applied in electromagnetism [89], consists in using homogenization methods which describes the rough surface by an equivalent transmission condition. We propose to apply it to the case of elastodynamic equations written as a first-order system. In particular, it would be very interesting to investigate if the rigorous techniques that have been used in [70], [71] can be applied to the theory of elasticity. This type of investigations could be a way for MAGIQUE-3D to consider medical applications where rough interfaces are often involved. Indeed, we would like to work on the modeling and the numerical simulation of ultrasonic propagation and its interaction with partially contacting interfaces, for instance bone/titanium in the context of an application to dentures, in collaboration with G. Haiat (University of Paris 7).

- **Asymptotic modeling**.

  In the context of wave propagation problems, we are investigating physical problems which involves multiple scales. Due to the presence of boundary layers (and/or thin layers, rough interfaces, geometric singularities), the direct numerical simulation (DNS) of these phenomenas involves a large numbers of degrees of freedom and high performance computing is required. The aim of this work is to develop credible alternatives to the DNS approach. Performing a multi-scale asymptotic analysis, we derive approximate models whose solution can be computed for a low computational cost. We study these approximate models mathematically (well-posedness, uniform error estimates) and numerically (we compare the solution of these approximate models to the solution of the initial model computed with high performance computating).

  We are mostly interested in the following problems.

  – 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;

  – ultrasonic wave propagation through bone-titanium media in medicine in collaboration with Inria Team MC2, and MSME;

  – asymptotic modeling of multi perforate plates in turbo reactors in collaboration with Cerfacs, INSA-Toulouse, Onera and Snecma in the framework of the ANR APAM.

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 calculus can be performed locally on each element of the mesh. The communications are ensured 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 to 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 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.** We are now planning to port the code on the new Intel Many Integrated Core Architecture (Intel MIC). The optimization of this code should begin in 2013, in collaboration with Dider Rémy of SGI.

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 impossible to reach in the past.

MOISE Project-Team

3. Scientific Foundations

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. Scientific Foundations

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.

STEEP Exploratory Action

3. Scientific Foundations

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 [2], 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 [3]. 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.

[2] http://www.modelistica.com/english
[3] 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. 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 cover model (CLUE-S [4]) which belongs to the last category.

In the following, our three main research axes are described.

3.2. Model calibration and validation

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

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

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

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

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

3.3. Sensitivity analysis

[4] http://www.ivm.vu.nl/en/Organisation/departments/spatial-analysis-decision-support/Clue

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

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

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

3.4. Modeling of socio-economic and environmental interactions

Considering the assessment of socio-economic impacts on the environment and ecosystem service analysis, the problems encountered here are intrinsically interdisciplinary: they draw on social sciences, ecology or Earth sciences. The modeling of the considered phenomena must take into account many factors of different nature which interact *via* various 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. The spatial processes involve proximity relationships and neighborhoods, like for example, between two adjacent parcels of land. 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. The multi-scale approaches can also be justified by the lack of data at the relevant scales. This is for example the case for the material flow analysis at local scales for which complex data disaggregations are required.

At this stage, it is crucial to understand that the scientific fields considered here are far from being mature. For example, the very notions of ecosystem services or local ecological accounting are quite recent and at best partially documented, but advances in those fields are essential, and will be required to identify transition paths to sustainability. Nowadays, the analyses are only qualitative or statistic. The phenomena are little understood.

Our goal here is then to do upstream research. It is to anticipate and to help the development of modeling tools that will be used tomorrow in these fields.

Developing 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; cellular automata, multi-agent models, system dynamics, or large systems of equations describing equilibrium models? 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? How to get models which automatically adapt to the granularity of the data and which are always numerically stable? 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?
Providing satisfying answers to these questions is a long term goal for STEEP.

BANG Project-Team

3. Scientific Foundations

3.1. Introduction

The dynamics of complex physical or biophysical phenomena involving many particles, including biological cells - which can be seen as active particles -, can be represented efficiently either by explicitly considering the behaviour of each particle individually or by Partial Differential Equations which, under certain hypotheses, represent local averages over a sufficiently large number of particles.

Since the XIXth century this formalism has shown its efficiency and ability to explain both qualitative and quantitative behaviours. The knowledge that has been gathered on such physical models, on algorithms for solving them on computers, on industrial implementation, opens the hope for success when dealing with life sciences also. This is one of the main goals of BANG. At small spatial scales, or at spatial scales of individual matter components where heterogeneities in the medium occur, agent-based models are developed. They complement the partial differential equation models considered on scales at which averages over the individual components behave sufficiently smoothly.

3.2. Mathematical modelling

What are the relevant physical or biological variables, what are the possible dominant effects ruling their dynamics, how to analyse the information coming out from a mathematical model and interpret them in the real situations under consideration ? These are the questions leading to select a mathematical model, generally also to couple several of them in order to render all physical or biomedical features which are selected by specialist partners (engineers, physicists, medical doctors). These are usually based on Navier-Stokes system for fluids (as in free surface fluid flows), on parabolic-hyperbolic equations (Saint-Venant system for shallow water, flows of electrons/holes in semiconductors, Keller-Segel model of chemotaxis).

3.3. Multiscale analysis

The complete physical or biomedical description is usually complex and requires very small scales. Efficiency of computer resolution leads to simplifications using averages of quantities. Methods allowing to achieve that goal are numerous and mathematically deep. Some examples studied in BANG are

- Coupled multiscale modelling (description of tumours and tissues from the sub-cellular level to the organ scale).
- Description of cell movement from the individual to the collective scales.
- Reduction of full 3d Navier-Stokes system to 2d or 1d hyperbolic equations by a section average (derivation of Saint-Venant system for shallow water).

3.4. Numerical Algorithms

Various numerical methods are used in BANG. They may be based on finite elements or finite volume methods, or stochastic methods for individual agents. Algorithmic improvements are needed in order to take into account the specificity of each model, of their coupling, or their 3D features. Among them we can mention

- Well-balanced schemes for shallow water system.
- Free-surface Navier-Stokes solvers based on a multilayer St-Venant approach.
- Deterministic and stochastic agent based models for the simulation of multi-cellular systems.

3.5. Proliferation dynamics and its control

- Cell division cycle in structured cell populations.
- Physiological and pharmacological control of cell proliferation.
- Physical mechanisms and constraints in cell proliferation.
- Intracellular spatiotemporal dynamics of genes and proteins: p53.
- Cell darwinism and drug resistance in cancer cells.
- Optimisation of cancer chemotherapy.
- Protein polymerisation and application to amyloid diseases.
- Inverse Problem for growth-fragmentation equations.

3.6. Tissue growth, regeneration and cell movements

This research activity aims at studying mathematical models related to tumour development and tissue organisation. Among the many biological aspects, examples are:

- Biomedical aspects of cell-cell interactions at the local and whole organ level.
- Migration of cells in tissues.
- Growth control of living tissues and organs.
- Regenerative medicine.
- Early embryology, and biomechanical aspects of cell interaction.
- Chemotaxis, self-organisation in cell populations.

3.7. Neurosciences

Cortical networks are constituted of a large number of statistically similar neurons in interaction. Each neuron has a nonlinear dynamics and is subject to noise. Moreover, neurological treatment involve several timescales. Multiscale analysis, both in spatial (number of cells) and temporal hence also constitute mathematical foundations of our approaches to neurosciences. In addition to the techniques described in section 3.1 - 3.4, our approach of the activity of large cortical areas involve:

- limit theorems of stochastic interacting particles systems, such as coupling methods or large deviations techniques, as used in mathematical approaches to the statistical physics of gases
- bifurcation analysis of deterministic and stochastic differential equation used to analyze the qualitative behaviour of networks
- singular perturbation theory, geometrical and topological approaches in dynamical systems used to uncover the dynamics in the presence of multiple timescales.

3.8. Free surface flows

Several industrial applications require to solve fluid flows with a free surface. BANG develops algorithms in two directions. Firstly flows in rivers and coastal areas using the Saint-Venant model with applications to dam break and pollution problems in averaged shallow water systems. Secondly, 3D hydrostatic flows by a multilayer Saint-Venant approach and 3D Navier-Stokes flows.

BIGS Project-Team

3. Scientific Foundations

3.1. Online data analysis

Participants: J-M. Monnez, R. Bar, 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 [61], 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 [58]), which allows to approximate eigenvectors in a stepwise manner. A systematic study of Principal Component and Factorial Analysis has then been leaded by Monnez in the series of papers [64], [62], [63], 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 ϵ. 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. [59]). Procedures testing the normality of the ϵ_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 [46], [47], we address the following questions:

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

We also obtain sharp estimates on the conditional distribution function in [48].

3.3. Stochastic modeling for complex and biological systems

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 [31] as well as mixed effects for parameter estimation.

• Another important aspect to quantify for PDT 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 leaded us to introduce a new model based on multi-state Markov chains arguments [10], 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 [39]. 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 [65], [68]) 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 [42], 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 [60], but also invoking the Russo-Vallois integration techniques [67]. The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in [43].

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 medecine (see [38]). Moreover, the fractional Brownian motion cannot be used to model some phenomena for which the regularity varies with time. Hence, many generalization (gaussian or not) of this model has been recently proposed, see for instance [32] for some Gaussian locally self-similar fields, [54] for some non-Gaussian models, [36] for anisotropic models.

Our team has thus contributed [41], [55], [54], [56], [66] and still contributes [35], [37], [36], [57], [49] 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

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 [72] 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 [33] 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. (i) 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 [31], 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.monolix.org) developed 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. (ii) 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 [44]. 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. (iii) 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 [53] 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 [45].

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 [30], [40]. 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 [69], dealing with some very particular families of equations, which do not cover the cases of interest for us.

BIOCORE Project-Team

3. Scientific Foundations

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 the living 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 being irrelevant to the problem at hand, the uncertainties and noise in experiments or even in the biological interactions require the development of 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 the living, 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, be it 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 [7]. 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. The interplay and information transfer between the different levels and scales within a biosystem/ecosystem introduces many new dynamical aspects. 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 to optimize some production or behavior at higher level: for example, control the growth of microalgae via their genetic or metabolic networks, to optimize the production of lipids for bioenergy at the photobioreactor level.

CARMEN Team

3. Scientific Foundations

3.1. Complex models for the propagation of cardiac action potentials

Cardiac arrythmias 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 developped, based on previous work [11], [3], [12] 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 developped. Both problems involve solving inverse problems that cannot be addressed with the more compex models. We want both to develop simple and fast models of the propagation of cardiac action potentials and improve the solutions to the inverse problems found in cardiac electrical imaging techniques.

The cardiac inverse problem consists in finding the cardiac activation maps or, more generally the whole cardiac electrical activity, from high density body surface electrocardiograms. It is a new and a powerful diagnosis technique, which success would be considered as a breakthrough in the cardiac diagnosis. Although widely studied during the last years, it remains a challenge for the scientific community. In many cases the quality of reconstructed electrical potential is not sufficiently accurate. The methods used consist in solving the Laplace equation on the volume delimited by the body surface and the epicardial surface.We 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 be using new regularization strategies and the theory of optimal control, both in the quasistatic and in the dynamic contexts.

Of cours we will use our models as a basis to regularize these inverse problems. We will conside the follwong strategies:

- using complete propagation models in the inverse problem, like the bidomain equations; for instance in order to localize 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.

Additionnaly, we will need to develop numerical techniques dedicated to our simplified eikonal/levl-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 [13], [14] and [19] and [2], we will focus on

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

DRACULA Project-Team

3. Scientific Foundations

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).

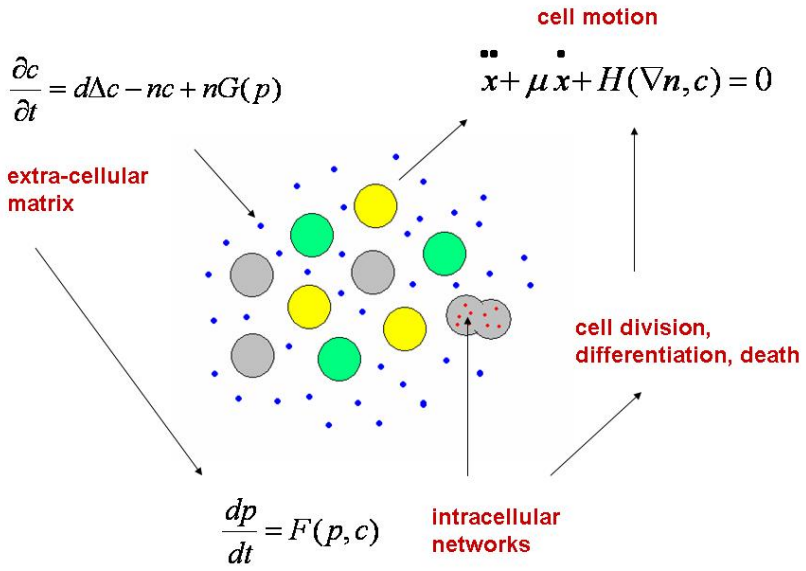


Figure 1. Schema of multi-scale models of cell dynamics: DPD-PDE-ODE models.

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 [35] 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 [36] 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 ([41], [43]), and they proved effective in hematopoiesis modelling ([30], [31], [33]). 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 [43]. The nonlocality can be either in the structure variable or in the time variable [30]. 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 [37].

3.5. Stochastic Equations

How identical cells perform different tasks may depend on deterministic factors, like external signals or pre-programming, or on stochastic factors. Intra-cellular processes are inherently noisy due to low numbers of molecules, complex interactions, limited number of DNA binding sites, the dynamical nature of molecular interactions, etc. Yet at the population level, deterministic and stochastic systems can behave the same way because of averaging over the entire population. This is why it is important to understand the causes and the roles of stochasticity in intra-cellular processes. In its simplest form, stochastic modelling of gene regulation networks considers the evolution of a low number of molecules (integer number) as they are synthesized, bound to other molecules, or degraded. The number n(t) of molecules at time t is a stochastic process whose probability transition to n+1 or n-1 is governed by a specific law. In some cases, master equations can yield analytical solutions for the probability distribution of n, P(n(t)). Numerically, efficient algorithms have been developed (Gillespie algorithms and variants) to handle statistically exact solutions of biochemical reactions. Recently, these algorithms have been adapted to take into account time delays. This allows a stochastic description of delayed regulatory feedback loops, both at the intra-cellular and the population levels. Another approach with stochastic differential equation, using Langevin equations is relevant to study extrinsic sources of noise on a system. A thesis (R. Yvinec) supervised by L. Pujo-Menjouet and M.C. Mackey devoted to "stochastic differential equations", started in Lyon on October 2009.

MACS Project-Team

3. Scientific Foundations

3.1. Formulation and analysis of effective and reliable shell elements

Thin structures (beams, plates, shells...) are widely considered in engineering applications. However, most experts agree that the corresponding discretization procedures (finite elements) are not yet sufficiently reliable, in particular as regards shell structures. A major cause of these difficulties lies in the numerical locking phenomena that arise in such formulations [2].

The expertise of the team in this area is internationally well-recognized, both in the mathematical and engineering communities. In particular, we have strongly contributed in analysing – and better explaining – the complex locking phenomena that arise in shell formulations [2]. In addition, we have proposed the first (and only to date) shell finite element procedure that circumvents locking [6]. However, the specific treatment applied to avoid locking in this procedure make it unable to correctly represent membrane-dominated behaviors of structures (namely, when locking is not to be expected). In fact, a "perfect shell element" – namely, with the desired reliability properties mathematically substantiated in a general framework – is still to be discovered, whereas numerous teams work on this issue throughout the world.

Another important (and related) issue that is considered in the team pertains to the design and analysis of numerical procedures that are adapted to industrial applications, i.e. that fulfill some actual industrial specifications. In particular, in the past we have achieved the first mathematical analysis of "general shell elements" – which are based on 3D variational formulations instead of shell models – these elements being among the most widely used and most effective shell elements in engineering practice.

3.2. Stability and control of structures

Stability of structures is – of course – a major concern for designers, in particular to ensure that a structure will not undergo poorly damped (or even unbounded) vibrations. In order to obtain improved stability properties – or to reach nominal specifications with a thinner a lighter design – a control device (whether active, semi-active, or passive) may be used.

The research performed in the team in this area – other than some prospective work on robust control – has been so far primarily focused on the stability of structures interacting with fluid flows. This problem has important applications e.g. in aeronautics (flutter of airplane wings), in civil engineering where the design of long-span bridges is now partly governed by wind effects, and in biomechanics (blood flows in arteries, for instance). Very roughly, the coupling between the structure and the flow can be described as follows: the structural displacements modify the geometry of the fluid domain, hence the fluid flow itself which in turn exerts an action on the structure. The effects of structural displacements on the fluid can be taken into account using ALE techniques, but the corresponding direct simulations are highly CPU-intensive, which makes stability analyses of such coupled problems very costly from a computational point of view. In this context a major objective of our work has been to formulate a simplified model of the fluid-structure interaction problem in order to allow computational assessments of stability at a reasonable cost.

3.3. Modeling and estimation in biomechanics

A keen interest in questions arising from the need to model biomechanical systems – and to discretize such problems – has always been present in the team since its creation. Our work in this field until now has been more specifically focused on the objectives related to our participation in the ICEMA ARC projects and in the CardioSense3D initiative, namely, to formulate a complete continuum mechanics model of a beating heart, and to confront – or "couple", in the terminology of the Inria strategic plan – numerical simulations of the model with actual clinical data via a data assimilation procedure.

Our global approach in this framework thus aims at using measurements of the cardiac activity in order to identify the parameters and state of a global electromechanical heart model, hence to give access to quantities of interest for diagnosing electrical activation and mechanical contraction symptoms. The model we propose is based on a chemically-controlled constitutive law of cardiac myofibre mechanics consistent with the behavior of myosin molecular motors [27]. The resulting sarcomere dynamics is in agreement with the "sliding filament hypothesis" introduced by Huxley. This constitutive law has an electrical quantity as an input which can be independently modeled, considered as given (or measured) data, or as a parameter to be estimated.

MASAIE Project-Team

3. Scientific Foundations

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, allow 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 [33]. 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 [27], [29]. 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) will be 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 is 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

$$\begin{cases} \dot{x} = f(x,u) \\ y = h(x), \end{cases} \tag{13}$$

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 the system (1) is a dynamical system

$$\dot{\widehat{x}}(t) = g(\widehat{x}(t), y(t), u(t)), \tag{14}$$

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

$$\|x(t) - \widehat{x}(t)\| \leq c\, \|x(0) - \widehat{x}(0)\|\, e^{-a\,t}\ ,\ \ \forall t > 0.$$

or at least $\|x(t) - \widehat{x}(t)\|$ 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. Gouze, O. Bernard, G. Sallet) and is still very active in this area. Now the problem is to identify relevant class 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 there is only one case of use of an observer in virology (Velasco-Hernandez J. , Garcia J. and Kirschner D. [38]) in modeling the chemotherapy of HIV, but this observer, based on classical linear theory, 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 becomes a near certainty. The rationale for studying continuous delays are 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 [5].

MODEMIC Project-Team

3. Scientific Foundations

3.1. Modelling and simulating microbial ecosystems

Microbial ecosystems naturally put into play phenomena at different scales, from the individual level at a microscopic scale to the population level at a macroscopic scale, with sometimes intermediate levels. The size of substrate molecules is a thousand time smaller than the size of microorganisms and usually diffuse much faster. The substrate consumption of one microorganism is negligible at the population level but the sum of the consumption of its neighbors can modify the local concentration of substrate, which itself modifies microorganism growth, acting as a *feedback loop*. For other variables that change slowly (pH, temperature...) cumulative effects create intermediate time scales, coupling individual and environment dynamics. The very large populations justify macroscopic modelling but for some ecosystems, spatial structures seen at intermediate scale need to be tackled. This is typically the case of biofilm ecosystems, for which the biofilm structure is responsible of characteristics of the overall ecosystem. Models that are purely individual-based or purely populational are rarely truly satisfactory to incorporate current knowledge on microbial ecosystems at various scales and to push ahead mathematical analysis or to derive operational rules.

3.1.1. Macroscopic models

The starting point is the knowledge of biologists that report a large number of mechanisms discovered or shown on laboratory experiments at a population level, such as competition for a growth-limiting substrate, predation interactions, obligate mutualism or communication between bacteria. If each *elementary* mechanism is today well understood and modelled at a macroscopic level, the consideration of several mechanisms together in a single model is still raising several questions of understanding and prediction. This is typically the case when there is more than one growth-limiting substrate in the chemostat model or when one couples species competition with a spatial structure (flocculation, niches...).

1. *Non-spatial models.*
   Ordinary differential equations (ODE) are the common way to describe the evolution of the size or concentration of species populations and their functional contribution in resource transformation (such as substrate degradation) in homogeneous or perfectly mixed compartments (or ecological niches). The well-known chemostat model [87] used in microbiology for single strain:

$$\begin{aligned} \dot{s} &= -\tfrac{1}{y}\mu(s)\,b + D\,(s_{in} - s) \\ \dot{b} &= \mu(s)\,b - D\,b \end{aligned}$$

   (where s and b stand respectively for the substrate and biomass concentrations, $\mu(s)$ is specific growth function, D the dilution rate, s_{in} the input substrate concentration), has to be extended to cope with the specificity of microbial ecosystems in the following directions.

   - very large number (hundreds or thousands) of species. This leads to the problem of characterization of their distribution during the transients, that is a way to study the *functional redundancy* of ecosystems.
   - environmental fluctuations (input flow rate, input concentration, temperature, pH...). This impacts the efficiency of a microbial ecosystem, when biological and environmental time scales are different. *Singular perturbations* is the technique we use to separate *slow* variables from *fast* ones, leading to approximations of the dynamics on *slow manifolds* to be determined and analyzed.

- interactions due to several limited resources and trophic chains. Most of the literature on the chemostat considers models with single limited resource, while some work studied purely essential or substitutable resources.
- several populations of bacteria (for each species) to describe the effects of certain spatial structures that are artificially created in bioreactors or naturally found in soils, like flocks, colonies or biofilms: the planktonic (or free) cells and the biofilm (or fixed) biomass (for telluric ecosystems, such a distinction is also relevant to represent the sticking/non sticking characteristics of soil). Considering simple models of aggregates (that are not spatialized) can provide a simplified model of the dynamics of the overall biomass.
- *active* and *dormant* bacteria. This distinction is motivated by the observations made on ecosystems of sparse resources such as arid soils.

2. *Spatial models.*
   In the spirit of lattice differential equations, representations in terms of networks of (abstract) interconnected bioreactors propose an intermediate level between models of average biomass (a single ODE) and a continuous representation of space (PDE). A model of interconnected bioreactors is a way to *implicitly* take into account spatial heterogeneity, without requiring a precise knowledge of it. It is similar to the island models used in ecology [85] but coupled with the dynamics of abiotic resources and hydrodynamics laws (transport, percolation, diffusion) governing the transfers between patches. This approach appears to be relevant for telluric ecosystems, for which pedologists report that microbial activities in soil are usually concentrated in *hot-spots* that could be seen as small bioreactors. Understanding the role of the topology of the interconnection network and how a spatial structure impacts the outputs is also relevant in biotechnology to improve the yield or stability of processes.

3.1.2. Microscopic models

In these models (birth and death processes, neutral models, individual-based models) the dynamic of the population is described in terms of discrete events: birth and death of individuals, or jumps in terms of biomass. These models can be gathered under the same framework that could be called *Markov stochastic processes with discrete events*. Most of the time they should be coupled with continuous components like the size of each individual or the dynamic of the resources (represented in terms of ODE or PDE).

The Markovian framework allows on the one hand sharp analyses and rescaling techniques [76]; on the other hand it induces a simplification in the memory structure of the process that is important in terms of simulation. Indeed, as the future state of the system depends from the past only through the present state, only the current state should be kept in memory for simulation.

We will consider three families of processes with discrete events, from simplest to most complex.

1. *Birth and death processes.*
   These models [77] are of first importance in small population size. They indeed allow investigation of near-to-extinction situations in a more realistic ways than the classical ODE models: they permit the computation, analytically but most of the time numerically, the distribution of extinction time and the probability of extinction. Efforts should be made to developed efficient Monte Carlo simulation procedures and approximation techniques for extinction probability and time distribution evaluation. In larger population sizes, they are advantageously approximated by diffusion models (see next section).
2. *The neutral models.*
   In *neutral* models [82] sizes of different species evolve as birth and death processes with immigration: all individuals have the same characteristics and are not spatialized. Such hypotheses could be considered unrealistic from a purely biological perspective, but these models focus on some precise properties to be simulated and predicted (for instance the biodiversity).

Comparing the prediction of species abundance of these models to real observations provides a way to justify or invalidate the neutral hypothesis. Extensions of the neutral model, that was originally introduced for forest ecology, have to be developed in order to better suit the framework of microbial ecology, such as the non constant size of the populations and spatialized variations.

3. *The individual based models.*
   IBM's [68], [69] appear to be well suited to describe colonies or biofilms [88]: in addition to birth, death and movement events, one has to consider *aggregation* and *detachment* events. The mechanisms that lead to the emergence of spatial patterns of colonies, or the formation of biofilms, which adhere to surface via polymers generated by the bacteria under specific hydrodynamics conditions, are not well understood yet. Typically, one can consider that bacteria inside the aggregates are disadvantaged to access the nutrient.

IBM modelling is a convenient way to propose aggregation and detachment mechanisms at the individual level in terms of random events connected to the geometry of the neighborhood, and to compare generated images with microscopic observations (for instance the confocal microscopy).

One has to be aware that few methods are available to study systematically and rigorously the properties of IBM, contrary to models based on differential equations (ODE, PDE...).

3.1.3. Bridges between models

The "theory of a computational model", that combine two kinds of models (typically ODE and IBM) that are different representations of the same objects, relies on two steps: the "program making" and the "theoretical study", in the spirit of the *double modelling* approach (roughly speaking, it consists in grasping the complexity of a IBM by analyzing accurately the consequences of each hypothesis on the macroscopic behavior of the model, building an approximate model of its global dynamics). Two main tools can be considered.

1. *Change of scale.*
   For IBM models (neutral or Markovian), we consider mean field and moments approximation techniques that provide information at the macroscopic (i.e. populational) level, to be compared with macroscopic models. From a birth-and-death process describing the individual level, a renormalisation can provide a stochastic differential equation at a meso-scale. The *diffusion approximation* technique can be understood as a numerical acceleration technique where the number of births and deaths follows a normal law. These stochastic models at meso-scale can provide additional information compared to deterministic models at a macro-scale, such as parameter identifiability or finite time extinction. The price to pay is to give much more conceptual and numerical efforts, that become less relevant for very large populations.

   For PDE models on spatial domains described with regular patterns (such as models of biofilm), the homogenization technique allows to obtain simpler PDE with constant parameters.

2. *The multi-scale modelling.*
   The spatial heterogeneity in microbial ecosystems requires to consider simultaneously several scales:
   - a *physical* scale. In batch processes, nutrient diffusion can be modelled by adapting the heat equation with Dirichlet boundary conditions. In continuous reactors, a convection-diffusion equation with Neumann boundary conditions is considered instead, the speed vector field being provided by the equations of fluid mechanics. The spatial scale used for the discretization is given by diffusion and flow parameters.
   - a *biological* scale, given by the size and mobility of bacteria. Usually, this scale is larger than the physical one (at least in the liquid phase).
   - an *aggregation* scale of colonies or biofilms, even larger, that provides the spatial patterns.

It is always possible to describe all the processes at the smaller common scale and then use a global representation, but this leads to extremely long computation times. The challenge is to manage these overlapping scales together and guarantee the stability of the numerical schemes. This is the goal of the *multi-scale* approaches [84]. For microbial ecosystems, it consists in

1. proposing new representations of the various scales of aggregation of bacteria in a model, taking into account the attachment-detachment processes determined by the local hydrodynamics conditions. Here, discussions with specialists of fluid mechanics are required.
2. coupling diversity models (e.g. models based on the neutral assumption) with spatial models (that reproduce the patterns observed on images of microscopy) to better understand the link biodiversity/structure.
3. introducing new *control* variables, considered as independent variables, each of them describing a proper scale. For this purpose, we investigate different techniques available to determine such variables: *mean-field* approximation, *singular perturbations*, *unification by limiting layers* or
4. *renormalising*, that aims at detecting invariants among models of different scales.

3.2. Interpreting and analyzing experimental observations

The validation of microbial models on data is rarely a straightforward task, because observations are most of the time not directly related to the variables of the models. Techniques such as abundance spectrum provided by molecular biology or confocal imagery are relatively recent in the field of microbial ecosystems. The signals provided by theses devices leave many research questions open in terms of data interpretation and experiments design. One can distinguish three kinds of key information that are needed at the basis of model assumptions:

- structure of the communities (i.e. who is present?),
- nature of interactions between species (competition, mutualism, syntrophism...),
- spatial structure of the ecosystems.

3.2.1. Assessment of community structures

Ecosystems biodiversity can be observed at different levels, depending on the kind of observations. One usually distinguish:

1. *The taxonomic diversity.* Several techniques developed by molecular biologists can gather information on the genetic structure of communities:
   - *sequencing of a given gene in the community.* The RNA 16S gene is often chosen to identify bacteria or Archeae.
   - *molecular fingerprints.* Some regions in the sequence of the RNA 16S gene encode faithfully the taxa species and can be amplified by PCR techniques.
   - *the sequencing of the overall genetic material of a community* (meta-genomic)

   All these techniques bring new problems of data interpretation to estimate in a robust manner the properties of communities. The signals are combinations of contributions of abundances from each taxon. For an ecosystem with a limited diversity, composed of known species, the signal allows to determine with no ambiguity the abundances. In natural ecosystems, the signal is more complex and it is hopeless to determine uniquely the taxa distribution.
2. *The functional diversity.* It is usually observed at a larger scale, measuring the performances of the overall ecosystem to convert organic matter. The taxonomic diversity does not usually provide such information (it is possible to study *functional genes* but this is much more difficult than studying the 16S one).

A convenient way to study the functional performance of microbial community dynamics is to grow the same microbial community on different substrate compositions, and monitor its performance on these different substrates. Neutral community models [82] provide a reference for what would happen if no functional differences are present in the community. The deviation of experimental observations from neutral model predictions can be considered as a measure of functional diversity.

Understanding the links between taxonomic and functional diversity is currently a tremendous research question in biology about genotype/phenotype links, that one can also find in the specific context of microbial ecosystems.

3.2.2. Characterization of the interactions

The role of biodiversity and its preservation in ecosystems are research questions currently largely open in ecology. The nature and number of interactions between bacterial populations are poorly known, and are most probably a key to understand biodiversity. In the classical chemostat model, inter-specific interactions are rarely considered. Also in theoretical ecology, interaction information is typically encoded in an *interaction matrix*, but the coupling with common abiotic resources and the stoichiometry is hardly addressed in the models.

The information provided by confocal microscopy is also a way to estimate the distance of interactions between microorganisms and substrates. This knowledge is not often documented although it is crucial for the construction of IBM.

3.2.3. Observation of spatial structures

Schematically, one can distinguish two origins of spatialization:

1. due the physics of the environment. In bioprocesses, this happens typically for large tank size (inducing *dead zones*) or sludge accumulation making the suspension closer from a porous medium than a liquid one. Numerical experimentation can be driven, coupling a solver of the equations of the fluid mechanics with microbiology equations. Then, the spatial distribution of the biomass can be observed and used to calibrate simpler models. Typically, a dead zone is modelled as a diffusive interconnection between two perfect (abstract) tanks.

   But the biotechnology industry aims at considering more sophisticated devices than simple tanks. For instance, the fluidized bed technique consists in creating a counter-current with oxygen bubbles for preventing the biomass to leave the rector. In more complex systems, such as soil ecosystems, it is difficult to obtain faithful simulations because the spatial structure is rarely known with accuracy. Nevertheless, local observations at the level of pores can be achieved, providing information for the construction of models.
2. due to the formation of aggregates (flocks, biofilms...) or biomass wall attachment. Patterns (from ten to a hundred micro-meters) can be observed with confocal microscopy.

   Spatial distribution of bacteria, shape of patterns and composition of the aggregates help to express hypotheses on individual behaviors. But quantification and variability of images provided by confocal microscopy are difficult. An open question is to determine the relevant morphological indicators that characterize aggregation and the formation of biofilms.

3.3. Identifying, controlling and optimizing bioprocesses

The dynamics of the microbial models possess specificities that do not allow the application of the popular methods of the theory of automatic control [71], such as linear control, feedback linearization or canonical forms.

- positivity constraints. State variables, as well as control inputs, have to stay non-negative (input flow pump cannot be reversed because of contamination issues).
- non-linearity. Several models have non-controllable or non-observable linearizations when inhibition effects are present (i.e. change of monotonicity in the growth curves).
- model and measurement uncertainties. In biology, it is rarely relevant to consider model uncertainties as additive Gaussian or finite energy signals.

3.3.1. Software sensors and identification

Sensors in biology are often poor and do not provide the measurements of all the state variables of the models: substrate, strain and product concentrations. In addition, measurements are often spoilt by errors. For instance optical density measurements give an indirect measure of the biomass, influenced by abiotic factors that share the same medium.

Analytical techniques are well suited to ODE models of small dimension, such as:

- guaranteed set-membership observers, when the system is non observable or in presence of unknown inputs,
- (non-linear) changes of coordinates, when the system is observable but not in a canonical form for the construction of observers with exponential convergence.

Software sensors can be also derived with the help of simulation based approaches like particle filtering techniques [75], [74]. This method is suited to diffusion models that approximate birth and death processes. Such softwares will allow us to investigate the different sources of randomness: demography, environment, but mainly imprecision of the sensors.

Similarly, identification techniques for constant parameters are based on sensor models as well as demography and environmental randomness models. In this case, Bayesian and non-Bayesian statistical techniques can be used [73], [70].

3.3.2. Bioprocess stabilization

In bioprocesses, the most efficient bacterial species at steady state are often inhibited by too large concentrations of substrates (this corresponds to assuming that the growth function $s \mapsto \mu(s)$ in the classical chemostat model is non-monotonic). This implies that the washout equilibrium (i.e. disappearance of the biomass) can be attractive, making the bioprocess bi-stable.

A common way to globally stabilize the dynamics toward the efficient equilibrium is to manipulate the dilution rate D [71]. But a diminution of the input flow rate for the stabilization requires to have enough room for an upstream storage, which is an expensive solution especially for developing countries that need to be equipped with new installations.

Alternative ways are proposed to stabilize bioprocesses without restricting the input flow rate:

- either by *physical means*, in terms of recirculation and bypass loops, or membranes as a selective way to keep bacteria and their aggregates inside the tank and improve its efficiency.
- either by *biological means*. The *biological control* consists in adding a small quantity another species with particular growth characteristics, that will help the other species to win the competition in the end.

3.3.3. Optimal control of bioreactors

The filling stage of bioreactors, or "fed-batch", is often time consuming because the quantity of initial biomass is small and consequently the population growth is slow. The minimal time is a typical criterion for designing a filling strategy, but the optimal feedback synthesis is non trivial and may present singular arcs when the growth function is non-monotonic [86].

Recent progress have been made in the consideration of

- multi-species in sequential reactors (having more than one strain makes significantly more difficult to analyze singular arcs because of the higher dimensions of the state space, and there is little literature on the subject),
- energy consumption of flow pumps and the value of byproducts of the biological reactions such as biogas in the criterion (instead of minimal time or as penalties). Recent concerns about sustainable development encourage engineers to look for compromises between those objectives under constraints on output concentrations.

3.3.4. Plant design and optimization

We distinguish two kind of setups:

1. *The industrial setup.* A research question, largely open today, is to identify networks of interconnections of bioreactors that are the most relevant for industrial applications in terms of the following objectives:
   - reasonably simple configurations (i.e. with a limited number of tanks and connections),
   - significant improvement of the residence time at steady state over single or simpler configurations, or shapes of the reservoirs such that the total volume required for a given desired conversion factor at steady state is reduced.
2. *The bioremediation setup.* Typically, the concentration of pollutant in a natural reservoir is solution of a transport-diffusion PDE, but the optimal control of the transport term is almost not studied in the literature.

NUMED Project-Team

3. Scientific Foundations

3.1. Multiscale modeling and computations

3.1.1. Spatial complexity: collective motion of cells

The collective motion of cells (bacteries on a gel or endothelial cells during angiogenesis) is a fascinating subject, that involves a combinaison of random walk and chemotaxis. The modeling of these problems is still active, since the pioneering works of Keller and Segel, and the mathematical study of the arising equations is a very active area of research.

Vincent Calvez focuses its effort on the following questions:

- Mathematical analysis of the Keller-Segel model

  [In collaboration with J.A. Carrillo and J. Rosado (UAB, Barcelona)]

  Following McCann 1997 and Otto 2001, we interpret the classical Keller-Segel system for chemotaxis as a gradient flow in the Wasserstein space. The free-energy functional turns out to be homogeneous. This viewpoint helps to understand better blow-up mechanisms, and to derive rates of convergence towards self-similar profiles. We investigate more precisely linear diffusion, porous medium diffusion and fast diffusion in competition with various interaction kernels.

  [In collaboration with N. Meunier (Paris 5) and R. Voituriez (Paris 6)]

  Another project consists in analyzing some variant of the Keller-Segel system when the chemoattractant is secreted at the boundary of the domain. This is motivated by modeling issues in cell polarization.

- Kinetic models for bacterial collective motion

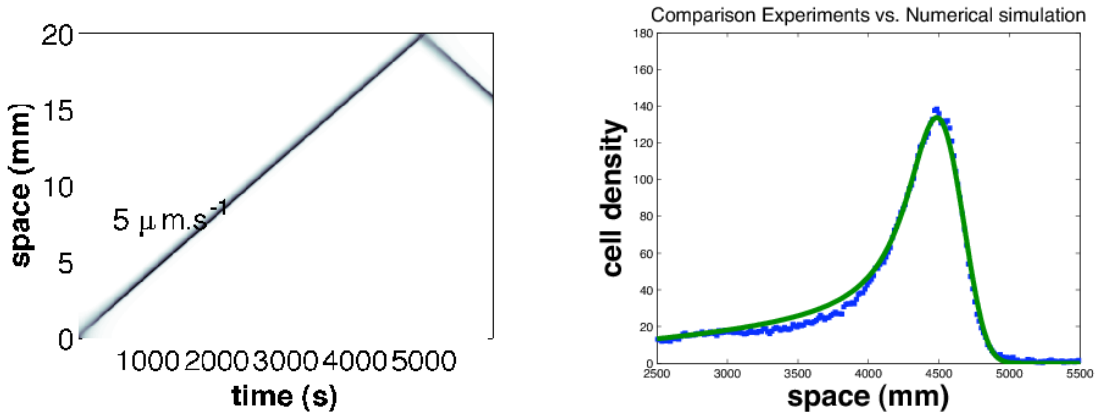


Figure 1. (left) Numerical simulation of a traveling pulse obtained with the kinetic model (right) Comparison between the bacteria density measured experimentally (blue dots) and the density computed from the kinetic model.

We have investigated kinetic models for bacterial chemotaxis following Alt and co-authors, Erban and Othmer, Dolak and Schmeiser.

We have developped a quantitative approach based on a couple of experiments performed by J. Saragosti in the team of A. Buguin and P. Silberzan (Institut Curie, Paris). These experiments describe with full statistical details solitary waves of bacteria E. coli in narrow channels. On the first set of experiments we have demonstrated that the drift-diffusion approximation of the kinetic model is valid and it fits the data very well (publication in PLoS Comput. Biol. 2010). On the second set of experiments we have simulated the kinetic model to obtain the best results as compared to the data (Fig. 1) (publication in PNAS 2011). Interestingly enough, the collaboration has led to the first experimental evidence of directional persistence of *E. coli* (the deviation angle after tumbling is smaller when the trajectory before tumbling goes in a favorable direction). We have demonstrated that this "microscopic effect" has a significant macroscopic influence on the solitary wave ($+30\%$ for the speed of the wave).

Based on these encouraging results, we have started a synthetic analysis of hyperbolic equations for chemotaxis and traveling waves.

In collaboration with Ch. Schmeiser (Univ. Vienna) we have investigated a simple (linear) kinetic equation for bacterial chemotaxis. We have obtained the existence of a stationary cluster (stable density distribution). We aim at applying the hypocoercivity results of Dolbeault-Mouhot-Schmeiser to derive a quantitative speed of relaxation towards the stable configuration. This work is under finalization.

In collaboration with N. Bournaveas (Univ. Edinburgh), C. di Russo (Univ. Lyon 1) and M. Ribot (Univ. Nice Sophia-Antipolis) we are studying hyperbolic models for cell motion. We improve the results obtained by Natalini-di Russo. These models are preliminary models which are to be complexified in order to describe growth of biofilms. This work is under progress.

In collaboration with E. Bouin and G. Nadin, we are analysing traveling waves arising in kinetic-growth equations. Namely, we study the coupling between a simple kinetic BGK operator (relaxation towards a given Maxwellian) and a logistic growth term. We have improved earlier results by Gallay-Raugel and Fedotov concerning the one-dimensional case with only two velocities. This work has been submitted. We continue the analysis with the full BGK operator. Counter-intuitive results have to be investigated further.

3.1.2. Modeling of spontaneous cell polarization

We have analysed recent models describing spontaneous polarization of cells (e.g. neuron growth cones or budding yeast). These models combine a diffusive term (in the cytoplasm) plus an advective field created at the membrane and diffusing in the cytoplasm (accounting for the actin network or the microtubules). This can be compared to the classical Keller-Segel model where diffusion competes with a non-local attractive field. Going beyond linear stability analysis we have used our know-how of the Keller-Segel system to derive global existence (no polarization) and blow-up (possibly polarization) criteria. We have also performed some numerical experiments to determine the models which exhibit spontaneous polarization. We have confirmed the prediction made by the physicists claiming that the microtubules cannot drive the cell into spontaneous polarization whereas the actin network can (Fig. 2).

Preliminaries results have been published in CRAS 2010 and SIAM J. Appl. Math (in press). We continue this project towards comparison with experimental data obtained in Matthieu Piel's lab at Institut Curie. A secondary goal consists in deriving a mechanistic model for the growth of the fission yeast Pombe. This is an ongoing work with A. Boudaoud (ENS de Lyon), N. Meunier (Univ. Paris 5), M. Piel (Institut Curie), P. Vigneaux (ENS de Lyon) and R. Voituriez (Univ. Paris 6). This is part of an ANR project JCJC, named "MODPOL" (Jan. 2012 – Dec. 2014). The project is coordinated by V. Calvez. It involves Th. Lepoutre (Inria Dracula), N. Meunier (Univ. Paris 5), M. Piel (Institut Curie), P. Vigneaux (ENS de Lyon) and R. Voituriez (Univ. Paris 6).

3.1.3. Polymerization-fragmentation processes

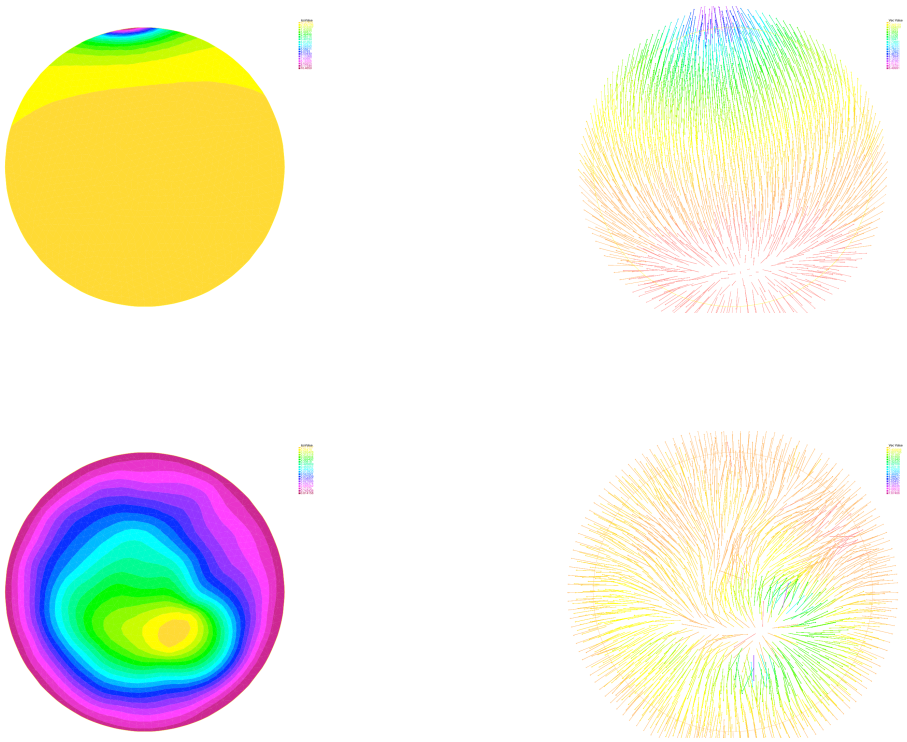


Figure 2. 2D numerical simulations of cell polarization on a round shaped cell. (Top) The actin network carries the attractive field: polarization occurs. (Bottom) The microtubules carry the attractive field: we observe no polarization. (Work in progress; simulations are done with FreeFEM++)

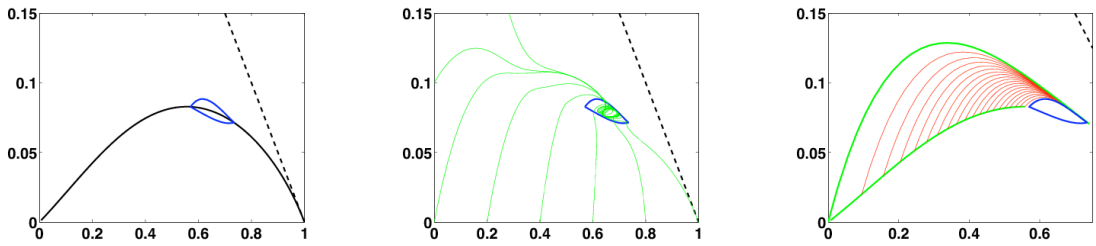


Figure 3. Dynamics of trajectories of the control system projected on the simplex. (left) Remarkable sets in the simplex: the line of eigenvectors parametrized by the control parameter, and the small "ergodic" set. (middle) All trajectories eventually enter the ergodic set. (right) We prove a tunelling effect: all trajectories are confined in a neighbourhood of the ergodic set, and moves towards it.

In collaboration with M. Doumic (Inria Bang) and P. Gabriel (Inria Beagle) we have studied the behaviour of the eigenvalue problem for genuine growth-fragmentation equations. We have focused on the dependence of the couple eigenvalue-eigenvector with respect to the growth and fragmentation coefficients. We have mainly used blowing-techniques and asymptotic estimates. We have shown counter-intuitive (non-monotonic) dependence. We have also discussed the possible consequences on applications.

Together with P. Gabriel (Inria Beagle) we are investigating the optimal control problem for a baby polymerization-fragmentation process mimicking the controlled growth of PrPres (prion) polymers. It consists in a three compartments system (small, intermediate and large polymers) with linear transitions between the compartments. We have a single control parameter acting on the fragmentation process.

We first assume that the control parameter has to be chosen constant. Under certain conditions, there is a best possible choice with infinite-time horizon. It maximizes the exponential growth by optimizing the eigenvalue of the polymerization-fragmentation matrix.

When we relax the condition of constant control, we have to deal with an optimal control problem. It can be translated into a Hamilton-Jacobi-Bellman equation. Although it is a very degenerated case, we can prove existence and uniqueness of an infinite-horizon eigenvalue, as in the constant case. We use the notion of ergodic set introduced by Arisawa-Lions (1998). The success of the proof relies on refined analysis of the dynamics of close-to-optimal trajectories projected on the simplex (Fig. 3). This work is under finalization.

3.1.4. Complex rheology

To investigate the growth of a tumor it is crucial to have a correct description of its mechanical aspects. Tumoral and normal cells may be seen as a complex fluid, with complex rheology.

Numerical investigations of complex flows is studied by P. Vigneaux who develops new numerical schemes fo Bingham type flows.

3.2. Parametrization of complex systems

The parametrization of complex systems in order to fit experimental results or to have a good qualitative behavior is a delicate issue since its require to simulate the complex systems for a large number of sets of parameters, which is very expensive.

In many medical contexts, the available data for one particular patient are rather poor (a few MRI for instance). However many patients are studied (20 to 100 or even more in frequent pathologies). Therefore it is difficult or even impossible to parametrize a model for a given patient (too many parameters with respect to the number of available clinical data). However, it is possible to infer the distribution of the parameters in the global population by using all the data of all the patients at the same time. This is the principle of populational parametrization: to look for the distribution of the parameters (Gaussian or log Gaussian) and not to try to study each patient individually.

Many algorithms have been developed for populational parametrization, in particular so called SAEM (Stochastic Approximation Expectation Maximization) algorithms, based on MCMC (Monte Carlo Markov Chain) algorithms. These algorithms are very expensive, and require hundreds of thousands of evaluations of the model. For ordinary differential equation based models, SAEM converge quickly (it takes ten to twenty minutes on a laptop for the Monolix implementation of SAEM. Monolix is developped by M. Lavielle at Inria).

However for PDE based models, the evaluation of one single model may be long (a few minutes, up to ten minutes), hence the evaluation of hundreds of thousands models is completely out of range. Moreover, SAEM can not be parallelized in an efficient way.

Numed has set a general strategy to allow populational approaches on complex systems or on PDE based models. It relies on a precomputation strategy, combined iteratively with SAEM algorithms.

With such a strategy, populational parametrization of a PDE like reaction diffusion equation (KPP) may be done on a few hours on a small cluster of cores (32 cores).

REO Project-Team

3. Scientific Foundations

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.

Because of the above mentioned difficulties, the interaction between the blood flow and the artery wall has often been neglected in most of the classical studies. The numerical properties of the fluid-structure coupling in blood flows are rather different from other classical fluid-structure problems. In particular, due to stability reasons it seems impossible to successfully apply the explicit coupling schemes used in 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 over the last few years several efficient fluid-structure interaction algorithms. We are now using these algorithms to address inverse problems in blood flows (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 [77] in the frame of combustion. It was later used to develop the Kiva code [67] at the Los Alamos National Laboratory, or the Hesione code [72], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [70], 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 [73]. 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 [1]. 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 plan to mix arbitrary Lagrangian Eulerian and fictitious domain approaches, or simplified valve models based on an immersed surface strategy.

[1] 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 [76], and Biot [68] 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 [69]), 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 to develop a multiscale modeling of the respiratory tract. Intraprenchymal airways distal from generation 7 of the tracheabronchial tree (TBT), which cannot be visualized by common medical imaging techniques, are modeled either by a single simple model or by a model set according to their order in TBT. The single model is based on straight pipe fully developed flow (Poiseuille flow in steady regimes) with given alveolar pressure at the end of each compartment. It will provide boundary conditions at the bronchial ends of 3D TBT reconstructed from imaging data. The model set includes three serial models. The generation down to the pulmonary lobule will be modeled by reduced basis elements. The lobular airways will be represented by a fractal homogenization approach. The alveoli, which are the gas exchange loci between blood and inhaled air, inflating during inspiration and deflating during expiration, will be described by multiphysics homogenization.

SISYPHE Project-Team

3. Scientific Foundations

3.1. System theory for systems modeled by ordinary differential equations

3.1.1. Identification, observation, control and diagnosis of linear and nonlinear systems

Characterizing and inferring properties and behaviors of objects or phenomena from observations using models is common to many research fields. For dynamical systems encountered in the domains of engineering and physiology, this is of practical importance for monitoring, prediction, and control. For such purposes, we consider most frequently, the following model of dynamical systems:

$$\begin{array}{rcl} \dfrac{dx(t)}{dt} & = & f(x(t),u(t),\theta,w(t)) \\ y(t) & = & g(x(t),u(t),\theta,v(t)) \end{array} \tag{15}$$

where $x(t)$, $u(t)$ and $y(t)$ represent respectively the state, input and output of the system, f and g characterize the state and output equations, parameterized by θ and subject to modeling and measurement uncertainties $w(t)$ and $v(t)$. Modeling is usually based on physical knowledge or on empirical experiences, strongly depending on the nature of the system. Typically only the input $u(t)$ and output $y(t)$ are directly observed by sensors. Inferring the parameters θ from available observations is known as system identification and may be useful for system monitoring [102], whereas algorithms for tracking the state trajectory $x(t)$ are called observers. The members of SISYPHE have gained important experiences in the modeling of some engineering systems and biomedical systems. The identification and observation of such systems often remain challenging because of strong nonlinearities [21]. Concerning control, robustness is an important issue, in particular to ensure various properties to all dynamical systems in some sets defined by uncertainties [83], [84]. The particularities of ensembles of connected dynamical systems raise new challenging problems.

Examples of reduced order models:
- Reduced order modeling of the cardiovascular system for signal & image processing or control applications. See section 3.3.1 .
- Excitable neuronal networks & control of the reproductive axis by the GnRH. See section 3.3.2 .
- Modeling, Control, Monitoring and Diagnosis of Depollution Systems. See section 6.1 .

3.2. System theory for quantum and quantum-like systems

3.2.1. Quantization of waves propagation in transmission-line networks & Inverse scattering

Linear stationary waves. Our main example of classical system that is interesting to see as a quantum-like system is the Telegrapher Equation, a model of transmission lines, possibly connected into a network. This is the standard model for electrical networks, where V and I are the voltage and intensity functions of z and k, the position and frequency and $R(z)$, $L(z)$, $C(z)$, $G(z)$ are the characteristics of the line:

$$\frac{\partial V(z,k)}{\partial z} = -\left(R(z)+jkL(z)\right)I(z,k), \quad \frac{\partial I(z,k)}{\partial z} = -(G(z)+jkC(z))V(z,k) \tag{16}$$

Since the work of Noordergraaf [101], this model is also used for hemodynamic networks with V and I respectively the blood pressure and flow in vessels considered as 1D media, and with $R=\frac{8\pi\eta}{S^2}$, $L=\frac{\rho}{S}$, $C=\frac{3S(r+h)}{E(2r+h)}$ where ρ and η are the density and viscosity of the blood ; r, h and E are the inner radius, thickness and Young modulus of the vessel. $S=\pi r^2$. The conductivity G is a small constant for blood flow.

Monitoring such networks is leading us to consider the following inverse problem: *get information on the functions R, L, C, G from the reflection coefficient* $\mathcal{R}(k)$ (ratio of reflected over direct waves) measured in some location by Time or Frequency Domain Reflectometry.

To study this problem it is convenient to use a Liouville transform, setting $x(z) = \int_0^z \sqrt{L(z')C(z')}dz'$, to introduce auxiliary functions $q^{\pm}(x) = \frac{1}{4}\frac{d}{x}\left(\ln \frac{L(x)}{C(x)}\right) \pm \frac{1}{2}\left(\frac{R(x)}{L(x)} - \frac{G(x)}{C(x)}\right)$ and $q_p(x) = \frac{1}{2}\left(\frac{R(x)}{L(x)} + \frac{G(x)}{C(x)}\right)$, so that (2) becomes a Zakharov-Shabat system [89] that reduces to a Schrödinger equation in the lossless case ($R = G = 0$):

$$\frac{\partial v_1}{\partial x} = (q_p - jk)\, v_1 + q^+ v_2, \quad \frac{\partial v_2}{\partial x} = -(q_p - jk)\, v_2 + q^- v_1 \tag{17}$$

and $I(x,k) = \frac{1}{\sqrt{2}}\left[\frac{C(x)}{L(x)}\right]^{\frac{1}{4}}(v_1(x,k) + v_2(x,k))$, $V(x,k) = -\frac{1}{\sqrt{2}}\left[\frac{L(x)}{C(x)}\right]^{\frac{1}{4}}(v_1(x,k) - v_2(x,k))$.
Our inverse problem becomes now an inverse scattering problem for a Zakharov-Shabat (or Schrödinger) equation: *find the potentials $q^{\pm}$ and q_p corresponding to* $\mathcal{R}$. This classical problem of mathematical physics can be solved using e.g. the Gelfand-Levitan-Marchenko method.

Nonlinear traveling waves. In some recent publications [92], [91], we use scattering theory to analyze a measured Arterial Blood Pressure (ABP) signal. Following a suggestion made in [103], a Korteweg-de Vries equation (KdV) is used as a physical model of the arterial flow during the pulse transit time. The signal analysis is based on the use of the Lax formalism: the iso-spectral property of the KdV flow allows to associate a constant spectrum to the non stationary signal. Let the non-dimensionalized KdV equation be

$$\frac{\partial y}{\partial t} - 6y\frac{\partial y}{\partial x} + \frac{\partial^3 y}{\partial x^3} = 0 \tag{18}$$

In the Lax formalism, y is associated to a Lax pair: a Schrödinger operator, $L(y) = -\frac{\partial^2}{\partial x^2} + y$ and an anti-Hermitian operator $M(y) = -4\frac{\partial^3}{\partial x^3} + 3y\frac{\partial}{\partial x} + 3\frac{\partial}{\partial x}y$. The signal y is playing here the role of the potential of $L(y)$ and is given by an operator equation equivalent to (4):

$$\frac{\partial L(y)}{\partial t} = [M(y), L(y)] \tag{19}$$

Scattering and inverse scattering transforms can be used to analyze y in term of the spectrum of $L(y)$ and conversely. The "bound states" of $L(y)$ are of particular interest: if $L(y)$ is solution of (5) and $L(y(t))$ has only bound states (no continuous spectrum), then this property is true at each time and y is a soliton of KdV. For example the arterial pulse pressure is close to a soliton [86], [14].

Inverse scattering as a generalized Fourier transform. For "pulse-shaped" signals y, meaning that $y \in L^1(\mathbb{R}; (1+|x|^2)dx)$, the squared eigenfunctions of $L(y)$ and their space derivatives are a basis in $L^1(\mathbb{R}; dx)$ (see e.g. [99]) and we use this property to analyze signals. Remark that the Fourier transform corresponds to using the basis associated with $L(0)$. The expression of a signal y in its associated basis is of particular interest. For a positive signal (as e.g. the arterial pressure), it is convenient to use $L(-y)$ as $-y$ is like a multi-well potential, and the Inverse scattering transform formula becomes:

$$y(x) = 4\sum_{n=1}^{n=N} \kappa_n \psi_n^2(x) - \frac{2i}{\pi}\int_{-\infty}^{-\infty} k\mathcal{R}(k) f^2(k,x)dk \tag{20}$$

where ψ_n and $f(k,.)$ are solutions of $L(-y)f = k^2 f$ with $k = i\kappa_n$, $\kappa_n > 0$, for ψ_n (bound states) and $k > 0$ for $f(k,.)$ (Jost solutions). The discrete part of this expression is easy to compute and provides useful informations on y in applications. The case $\mathcal{R} = 0$ ($-y$ is a reflectionless potential) is then of particular interest as $2N$ parameters are sufficient to represent the signal. We investigate in particular approximation of pulse-shaped signals by such potentials corresponding to N-solitons.

3.2.2. Identification & control of quantum systems

Interesting applications for quantum control have motivated seminal studies in such wide-ranging fields as chemistry, metrology, optical networking and computer science. In chemistry, the ability of coherent light to manipulate molecular systems at the quantum scale has been demonstrated both theoretically and experimentally [98]. In computer science, first generations of quantum logical gates (restrictive in fidelity) has been constructed using trapped ions controlled by laser fields (see e.g. the "Quantum Optics and Spectroscopy Group, Univ. Innsbruck"). All these advances and demands for more faithful algorithms for manipulating the quantum particles are driving the theoretical and experimental research towards the development of new control techniques adapted to these particular systems. A very restrictive property, particular to the quantum systems, is due to the destructive behavior of the measurement concept. One can not measure a quantum system without interfering and perturbing the system in a non-negligible manner.

Quantum decoherence (environmentally induced dissipations) is the main obstacle for improving the existing algorithms [88]. Two approaches can be considered for this aim: first, to consider more resistant systems with respect to this quantum decoherence and developing faithful methods to manipulate the system in the time constants where the decoherence can not show up (in particular one can not consider the back-action of the measurement tool on the system); second, to consider dissipative models where the decoherence is also included and to develop control designs that best confronts the dissipative effects.

In the first direction, we consider the Schrödinger equation where $\Psi(t,x)$, $-\frac{1}{2}\Delta$, V, μ and $u(t)$ respectively represent the wavefunction, the kinetic energy operator, the internal potential, the dipole moment and the laser amplitude (control field):

$$i\frac{d}{dt}\Psi(t,x) = (H_0 + u(t)H_1)\Psi(t,x) = (-\frac{1}{2}\Delta + V(x) + u(t)\mu(x))\Psi(t,x), \quad \Psi|_{t=0} = \Psi_0, \tag{21}$$

While the finite dimensional approximations ($\Psi(t) \in \mathbb{C}^N$) have been very well studied (see e.g. the works by H. Rabitz, G. Turinici, ...), the infinite dimensional case ($\Psi(t,.) \in L^2(\mathbb{R}^N;\mathbb{C})$) remains fairly open. Some partial results on the controllability and the control strategies for such kind of systems in particular test cases have already been provided [79], [80], [94]. As a first direction, in collaboration with K. Beauchard (CNRS, ENS Cachan) et J-M Coron (Paris-sud), we aim to extend the existing ideas to more general and interesting cases. We will consider in particular, the extension of the Lyapunov-based techniques developed in [95], [81], [94]. Some technical problems, like the pre-compactness of the trajectories in relevant functional spaces, seem to be the main obstacles in this direction.

In the second direction, one needs to consider dissipative models taking the decoherence phenomena into account. Such models can be presented in the density operator language. In fact, to the Schrödinger equation (7), one can associate an equation in the density operator language where $\rho = \Psi\Psi^*$ represents the projection operator on the wavefunction Ψ ($[A,B] = AB - BA$ is the commutator of the operators A and B):

$$\frac{d}{dt}\rho = -i[H_0 + u(t)H_1, \rho], \tag{22}$$

Whenever, we consider a quantum system in its environment with the quantum jumps induced by the vacuum fluctuations, we need to add the dissipative effect due to these stochastic jumps. Note that at this level, one also can consider a measurement tool as a part of the environment. The outputs being partial and not giving complete information about the state of the system (Heisenberg uncertainty principle), we consider a so-called quantum filtering equation in order to model the conditional evolution of the system. Whenever the measurement tool composes the only (or the only non-negligible) source of decoherence, this filter equation admits the following form:

$$\begin{aligned} d\rho_t = -i[H_0 + u(t)H_1, \rho_t]dt + (L\rho_t L^* - \frac{1}{2}L^*L\rho_t - \frac{1}{2}\rho_t L^* L)dt \\ +\sqrt{\eta}(L\rho_t + \rho_t L^* - \mathrm{Tr}\left[(L+L^*)\rho_t\right]\rho_t)dW_t, \end{aligned} \tag{23}$$

where L is the so-called Lindblad operator associated to the measurement, $0 < \eta \leq 1$ is the detector's efficiency and where the Wiener process W_t corresponds to the system output Y_t via the relation $dW_t = dY_t - \mathrm{Tr}\left[(L+L^*)\rho_t\right]dt$. This filter equation, initially introduced by Belavkin [82], is the quantum analogous of a Kushner-Stratonovic equation. In collaboration with H. Mabuchi and his co-workers (Physics department, Caltech), we would like to investigate the derivation and the stochastic control of such filtering equations for different settings coming from different experiments [96].

Finally, as a dual to the control problem, physicists and chemists are also interested in the parameter identification for these quantum systems. Observing different physical observables for different choices of the input $u(t)$, they hope to derive more precise information about the unknown parameters of the system being parts of the internal Hamiltonian or the dipole moment. In collaboration with C. Le Bris (Ecole des ponts and Inria), G. Turinici (Paris Dauphine and Inria), P. Rouchon (Ecole des Mines) and H. Rabitz (Chemistry department, Princeton), we would like to propose new methods coming from the systems theory and well-adapted to this particular context. A first theoretical identifiability result has been proposed [93]. Moreover, a first observer-based identification algorithm is under study.

3.3. Physiological & Clinical research topics

3.3.1. The cardiovascular system: a multiscale controlled system

Understanding the complex mechanisms involved in the cardiac pathological processes requires fundamental researches in molecular and cell biology, together with rigorous clinical evaluation protocols on the whole organ or system scales. Our objective is to contribute to to these researches by developing low-order models of the cardiac mechano-energetics and control mechanisms, for applications in model-based cardiovascular signal or image processing.

We consider intrinsic heart control mechanisms, ranging from the Starling and Treppe effects on the cell scale to the excitability of the cardiac tissue and to the control by the autonomous nervous system. They all contribute to the function of the heart in a coordinated manner that we want to analyze and assess. For this purpose, we study reduced-order models of the electro-mechanical activity of cardiac cells designed to be coupled with measures available on the organ scale (e.g. ECG and pressure signals). We study also the possibility to gain insight on the cell scale by using model-based multiscale signal processing techniques of long records of cardiovascular signals.

Here are some questions of this kind, we are considering:
- Modeling the controlled contraction/relaxation from molecular to tissue and organ scales.
- Direct and inverse modeling the electro-mechanical activity of the heart on the cell scale.
- Nonlinear spectral analysis of arterial blood pressure waveforms and application to clinical indexes.
- Modeling short-term and long-term control dynamics on the cardiovascular-system scale. Application to a Total Artificial Heart.

3.3.2. Reproductive system: follicular development & ovulation control

The ovulatory success is the main limiting factor of the whole reproductive process, so that a better understanding of ovulation control is needed both for clinical and zootechnical applications. It is necessary to improve the treatment of anovulatory infertility in women, as it can be by instance encountered in the PolyCystic Ovarian Syndrome (PCOS), whose prevalence among reproductive-age women has been estimated at up to 10%. In farm domestic species, embryo production following FSH stimulation (and subsequent insemination) enables to amplify the lineage of chosen females (via embryo transfer) and to preserve the genetic diversity (via embryo storage in cryobanks). The large variability in the individual responses to ovarian stimulation treatment hampers both their therapeutic and farming applications. Improving the knowledge upon the mechanisms underlying FSH control will help to improve the success of assisted reproductive technologies, hence to prevent ovarian failure or hyperstimulation syndrome in women and to manage ovulation rate and ovarian cycle chronology in farm species.

To control ovarian cycle and ovulation, we have to deeply understand the selection process of ovulatory follicles, the determinism of the species-specific ovulation rate and of its intra- and between-species variability, as well as the triggering of the ovulatory GnRH surge from hypothalamic neurons.
Beyond the strict scope of Reproductive Physiology, this understanding raises biological questions of general interest, especially in the fields of

Molecular and Cellular Biology. The granulosa cell, which is the primary target of FSH in ovarian follicles, is a remarkable cellular model to study the dynamical control of the transitions between the cellular states of quiescence, proliferation, differentiation, and apoptosis, as well as the adaptability of the response to the same extra-cellular signal according to the maturity level of the target cell. Moreover, the FSH receptor belongs to the seven transmembrane spanning receptor family, which represent the most frequent target (over 50%) amongst the therapeutic agents currently available. The study of FSH receptor-mediated signaling is thus not only susceptible to allow the identification of relaying controls to the control exerted by FSH, but it is also interesting from a more generic pharmacological viewpoint.

Neuroendocrinology and Chronobiology. The mechanisms underlying the GnRH ovulatory surge involve plasticity phenomena of both neuronal cell bodies and synaptic endings comparable to those occurring in cognitive processes. Many time scales are interlinked in ovulation control from the fastest time constants of neuronal activation (millisecond) to the circannual variations in ovarian cyclicity. The influence of daylength on ovarian activity is an interesting instance of a circannual rhythm driven by a circadian rhythm (melatonin secretion from the pineal gland).

Simulation and control of a multiscale conservation law for follicular cells

In the past years, we have designed a multiscale model of the selection process of ovulatory follicles, including the cellular, follicular and ovarian levels [11], [10]. The model results from the double structuration of the granulosa cell population according to the cell age (position within the cell cycle) and to the cell maturity (level of sensitivity towards hormonal control). In each ovarian follicle, the granulosa cell population is described by a density function whose changes are ruled by conservation laws. The multiscale structure arises from the formulation of a hierarchical control operating on the aging and maturation velocities as well on the source terms of the conservation law. The control is expressed from different momentums of the density leading to integro-differential expressions.
Future work will take place in the REGATE project and will consist in:
- predicting the selection outcome (mono-, poly-ovulation or anovulation / ovulation chronology) resulting from given combinations of parameters and corresponding to the subtle interplay between the different organs of the gonadotropic axis (hypothalamus, pituitary gland and ovaries). The systematic exploration of the situations engendered by the model calls for the improvement of the current implementation performances. The work will consist in improving the precision of the numerical scheme, in the framework of the finite volume method and to implement the improved scheme,

- solving the control problems associated with the model. Indeed, the physiological conditions for the triggering of ovulation, as well as the counting of ovulatory follicles amongst all follicles, define two nested and coupled reachability control problems. Such particularly awkward problems will first be tackled from a particle approximation of the density, in order to design appropriate control laws operating on the particles and allowing them to reach the target state sets.

Connectivity and dynamics of the FSH signaling network in granulosa cells

The project consists in analyzing the connectivity and dynamics of the FSH signaling network in the granulosa cells of ovarian follicles and embedding the network within the multiscale representation described above, from the molecular up to the organic level. We will examine the relative contributions of the Gα_s and βarrestin-dependent pathways in response to FSH signal, determine how each pathway controls downstream cascades and which mechanisms are involved in the transition between different cellular states (quiescence, proliferation, differentiation and apoptosis). On the experimental ground, we propose to develop an antibody microarray approach in order to simultaneously measure the phosphorylation levels of a large number of signaling intermediates in a single experiment. On the modeling ground, we will use the BIOCHAM (biochemical abstract machine) environment first at the boolean level, to formalize the network of interactions corresponding to the FSH-induced signaling events on the cellular scale. This network will then be enriched with kinetic information coming from experimental data, which will allow the use of the ordinary differential equation level of BIOCHAM. In order to find and fine-tune the structure of the network and the values of the kinetic parameters, model-checking techniques will permit a systematic comparison between the model behavior and the results of experiments. In the end, the cell-level model should be abstracted to a much simpler model that can be embedded into a multiscale one without losing its main characteristics.

Bifurcations in coupled neuronal oscillators.

We have proposed a mathematical model allowing for the alternating pulse and surge pattern of GnRH (Gonadotropin Releasing Hormone) secretion [5]. The model is based on the coupling between two systems running on different time scales. The faster system corresponds to the average activity of GnRH neurons, while the slower one corresponds to the average activity of regulatory neurons. The analysis of the slow/fast dynamics exhibited within and between both systems allows to explain the different patterns (slow oscillations, fast oscillations and periodical surge) of GnRH secretion.

This model will be used as a basis to understand the control exerted by ovarian steroids on GnRH secretion, in terms of amplitude, frequency and plateau length of oscillations and to discriminate a direct action (on the GnrH network) from an indirect action (on the regulatory network) of steroids. From a mathematical viewpoint, we have to fully understand the sequences of bifurcations corresponding to the different phases of GnRH secretion. This study will be derived from a 3D reduction of the original model.

VIRTUAL PLANTS Project-Team

3. Scientific Foundations

3.1. Analysis of structures resulting from meristem activity

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

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

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

3.2. Meristem functioning and development

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

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

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

OpenAlea is a component based, open-software platform for interdisciplinary research in plant modeling and simulation. This 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 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, LaBRI, Laboratory Jean Kuntzmann 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.).