RESEARCH CENTER

FIELD
Digital Health, Biology and Earth

Activity Report 2018

Section Scientific Foundations

Edition: 2019-03-07

ABS Project-Team

3. Research Program

3.1. Introduction

The research conducted by ABS focuses on three main directions in Computational Structural Biology (CSB), together with the associated methodological developments:
– Modeling interfaces and contacts,
– Modeling macro-molecular assemblies,
– Modeling the flexibility of macro-molecules,
– Algorithmic foundations.

3.2. Modeling interfaces and contacts

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

The Protein Data Bank, http://www.rcsb.org/pdb, contains the structural data which have been resolved experimentally. Most of the entries of the PDB feature isolated proteins [0], the remaining ones being protein - protein or protein - drug complexes. These structures feature what Nature does – up to the bias imposed by the experimental conditions inherent to structure elucidation, and are of special interest to investigate non-covalent contacts in biological complexes. More precisely, given two proteins defining a complex, interface atoms are defined as the atoms of one protein *interacting* with atoms of the second one. Understanding the structure of interfaces is central to understand biological complexes and thus the function of biological molecules [52]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [55]. Current investigations follow two routes. From the experimental perspective [37], 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 [49]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [44].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change [0], or a function stating that a complex is more stable than another one, in which case the value returned is a score and not an energy. Borrowing to statistical mechanics [31], 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 [53], [39]. 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 [45]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

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

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

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

In the static setting, one seeks the minimalist geometric model providing a relevant bio-physical signal. A first step in doing so consists of identifying interface atoms, so as to relate the geometry and the bio-chemistry at the interface level [12]. To elaborate at the atomic level, one seeks a structural alphabet encoding the spatial structure of proteins. At the side-chain and backbone level, an example of such alphabet is that of [32]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [54], 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 [36].

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 [48]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

3.3. Modeling macro-molecular assemblies

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

3.3.1. Reconstruction by Data Integration

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [30]. 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 [29], 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 [28], 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 [28]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

3.4. Modeling the flexibility of macro-molecules

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

Proteins in vivo vibrate at various frequencies: high frequencies correspond to small amplitude deformations of chemical bonds, while low frequencies characterize more global deformations. This flexibility contributes to the entropy thus the *free energy* of the system *protein - solvent*. From the experimental standpoint, NMR studies generate ensembles of conformations, called *conformers*, and so do molecular dynamics (MD) simulations. Of particular interest while investigating flexibility is the notion of correlated motion. Intuitively, when a protein is folded, all atomic movements must be correlated, a constraint which gets alleviated when the protein unfolds since the steric constraints get relaxed [0]. Understanding correlations is of special interest to predict the folding pathway that leads a protein towards its native state. A similar discussion holds for the case of partners within a complex, for example in the third step of the *diffusion - conformer selection - induced fit* complex formation model.

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

At the side-chain level, the question of improving rotamer libraries is still of interest [34]. 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) [51]. 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 [47], to Morse theory [42] and to analysis of meta-stable states of time series [43] have been proposed.

3.5. Algorithmic foundations

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

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

3.5.1. Modeling Interfaces and Contacts

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

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

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

3.5.2. Modeling Macro-molecular Assemblies

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

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

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

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

3.5.3. Modeling the Flexibility of Macro-molecules

Given a sampling on an energy landscape, a number of fundamental issues actually arise: how does the point cloud describe the topography of the energy landscape (a question reminiscent from Morse theory)? Can one infer the effective number of degrees of freedom of the system over the simulation, and is this number varying? Answers to these questions would be of major interest to refine our understanding of folding and docking, with applications to the prediction of structural properties. It should be noted in passing that such questions are probably related to modeling phase transitions in statistical physics where geometric and topological methods are being used [46].

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

BEAGLE Project-Team

3. Research Program

3.1. Introduction

As stated above, the research topics of the BEAGLE Team are centered on the modelization and simulation of cellular processes. More specifically, we focus on two specific processes that govern cell dynamics and behavior: Biophysics and Evolution. We are strongly engaged into the integration of these level of biological understanding.

3.2. Research axis 1: Computational cellular biochemistry

Biochemical kinetics developed as an extension of chemical kinetics in the early 20th century and inherited the main hypotheses underlying Van't Hoff's law of mass action : a perfectly-stirred homogeneous medium with deterministic kinetics. This classical view is however challenged by recent experimental results regarding both the movement and the metabolic fate of biomolecules. First, it is now known that the diffusive motion of many proteins in cellular media exhibits deviations from the ideal case of Brownian motion, in the form of position-dependent diffusion or anomalous diffusion, a hallmark of poorly mixing media. Second, several lines of evidence indicate that the metabolic fate of molecules in the organism not only depends on their chemical nature, but also on their spatial organisation – for example, the fate of dietary lipids depends on whether they are organized into many small or a few large droplets (see e.g. [36]). In this modern-day framework, cellular media appear as heterogeneous collections of contiguous spatial domains with different characteristics, thus providing spatial organization of the reactants. Moreover, the number of implicated reactants is often small enough that stochasticity cannot be ignored. To improve our understanding of intracellular biochemistry, we study spatiotemporal biochemical kinetics using computer simulations (particle-based spatially explicit stochastic simulations) and mathematical models (age-structured PDEs).

3.3. Research axis 2: Models for Molecular Evolution

We study the processes of genome evolution, with a focus on large-scale genomic events (rearrangements, duplications, transfers). We are interested in deciphering general laws which explain the organization of the genomes we observe today, as well as using the knowledge of these processes to reconstruct some aspects of the history of life. To do so, we construct mathematical models and apply them either in a "forward" way, *i.e.* observing the course of evolution from known ancestors and parameters, by simulation (*in silico experimental evolution*) or mathematical analysis (*theoretical biology*), or in a "backward" way, *i.e.* reconstructing ancestral states and parameters from known extant states (*phylogeny, comparative genomics*). Moreover we often mix the two approaches either by validating backwards reconstruction methods on forward simulations, or by using the forward method to test evolutionary hypotheses on biological data.

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

Brain cells are rarely considered by computational systems biologists, though they are especially well suited for the field: their major signaling pathways are well characterized, the cellular properties they support are well identified (e.g. synaptic plasticity) and eventually give rise to well known functions at the organ scale (learning, memory). Moreover, electro-physiology measurements provide us with an experimental monitoring of signaling at the single cell level (sometimes at the sub-cellular scale) with unrivaled temporal resolution (milliseconds) over durations up to an hour. In this research axis, we develop modeling approaches for systems biology of both neuronal cells and glial cells, in particular astrocytes. We are mostly interested in understanding how the pathways implicated in the signaling between neurons, astrocytes and neurons-astrocytes interactions implement and regulate synaptic plasticity.

3.5. Research axis 4: Evolutionary Systems Biology

This axis, consisting in integrating the two main biological levels we study, is a long-standing and long-term objective in the team. These last years we did not make significant advances in this direction and we even removed this objective from last year's report. However the evolution of the team staff and projects allows us to give it back its central place. We now have the forces and ideas to progress. We have several short and middle term projects to integrate biochemical data and evolution. In particular we are analysing with an evolutionary perspective the 3D conformation of chromosomes, the regulatory landscape of genomes, the chromatin-associated proteins.

BIGS Project-Team

3. Research Program

3.1. Introduction

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

3.2. Stochastic modeling

Our aim is to propose relevant stochastic frameworks for the modeling and the understanding of biological systems. The stochastic processes are particularly suitable for this purpose. Among them, Markov chains give a first framework for the modeling of population of cells [89], [66]. Piecewise deterministic processes are non diffusion processes also frequently used in the biological context [56], [65], [58]. Among Markov model, we developed strong expertise about processes derived from Brownian motion and Stochastic Differential Equations [81], [64]. For instance, knowledge about Brownian or random walk excursions [88], [80] helps to analyse genetic sequences and to develop inference about it. However, nature provides us with many examples of systems such that the observed signal has a given Hölder regularity, which does not correspond to the one we might expect from a system driven by ordinary Brownian motion. This situation is commonly handled by noisy equations driven by Gaussian processes such as fractional Brownian motion of fractional fields. The basic aspects of these differential equations are now well understood, mainly thanks to the so-called rough paths tools [72], but also invoking the Russo-Vallois integration techniques [82]. The specific issue of Volterra equations driven by fractional Brownian motion, which is central for the subdiffusion within proteins problem, is addressed in [57]. Many generalizations (Gaussian or not) of this model have been recently proposed for some Gaussian locally self-similar fields, or for some non-Gaussian models [69], or for anisotropic models [53].

3.3. Estimation and control for stochastic processes

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

There are many estimation techniques for diffusion processes or coefficients of fractional or multifractional Brownian motion according to a set of observations [68], [49], [55]. But, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. Our team has a good expertise about inference of the jump rate and the kernel of Piecewise Deterministic Markov Processes (PDMP) [45], [46], [44], [47]. However, there are many directions to go further into. For instance, previous works made the assumption of a complete observation of jumps and mode, that is unrealistic in practice. We tackle the problem of inference of "Hidden PDMP". As an example, in pharmacokinetics modeling inference, we want to take into account for presence of timing noise and identification from longitudinal data. We have expertise on this subjects [50], and we also used mixed models to estimate tumor growth [51].

We consider the control of stochastic processes within the framework of Markov Decision Processes [79] and their generalization known as multi-player stochastic games, with a particular focus on infinite-horizon problems. In this context, we are interested in the complexity analysis of standard algorithms, as well as the proposition and analysis of numerical approximate schemes for large problems in the spirit of [52]. Regarding complexity, a central topic of research is the analysis of the Policy Iteration algorithm, which has made significant progress in the last years [91], [78], [63], [87], but is still not fully understood. For large problems, we have a long experience of sensitivity analysis of approximate dynamic programming algorithms for Markov Decision Processes [85], [84], [86], [71], [83], and we currently investigate whether/how similar ideas may be adapted to multi-player stochastic games.

3.4. Algorithms and estimation for graph data

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

Network inference is the process of making inference about the link between two variables taking into account the information about other variables. [90] gives a very good introduction and many references about network inference and mining. Many methods are available to infer and test edges in Gaussian Graphical models [90], [73], [61], [62]. However, when dealing with abundance data, because inflated zero data, we are far from gaussian assumption and we want to develop inference in this case.

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

3.5. Regression and machine learning

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

Many methods are available to estimate conditional quantiles and test dependencies [77], [67]. Among them we have developed nonparametric estimation by local analysis via kernel methods [59], [60] and we want to study properties of this estimator in order to derive a measure of risk like confidence band and test. We study also many other regression models like survival analysis, spatio temporal models with covariates. Among the multiple regression models, we want to develop omnibus test that examine several assumptions together.

Concerning the analysis of high dimensional data, our view on the topic relies on the *French data analysis school*, specifically on Factorial Analysis tools. In this context, stochastic approximation is an essential tool [70], which allows one to approximate eigenvectors in a stepwise manner [76], [74], [75]. BIGS aims at performing accurate classification or clustering by taking advantage of the possibility of updating the information "online" using stochastic approximation algorithms [54]. We focus on several incremental procedures for regression and data analysis like linear and logistic regressions and PCA.

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

BONSAI Project-Team

3. Research Program

3.1. Sequence processing for Next Generation Sequencing

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

3.2. Noncoding RNA

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

3.3. Genome structures

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

3.4. Nonribosomal peptides

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

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

CAPSID Project-Team

3. Research Program

3.1. Classifying and Mining Protein Structures and Protein Interactions

3.1.1. Context

The scientific discovery process is very often based on cycles of measurement, classification, and generalisation. It is easy to argue that this is especially true in the biological sciences. The proteins that exist today represent the molecular product of some three billion years of evolution. Therefore, comparing protein sequences and structures is important for understanding their functional and evolutionary relationships [85], [57]. There is now overwhelming evidence that all living organisms and many biological processes share a common ancestry in the tree of life. Historically, much of bioinformatics research has focused on developing mathematical and statistical algorithms to process, analyse, annotate, and compare protein and DNA sequences because such sequences represent the primary form of information in biological systems. However, there is growing evidence that structure-based methods can help to predict networks of protein-protein interactions (PPIs) with greater accuracy than those which do not use structural evidence [61], [90]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [42].

3.1.2. Quantifying Structural Similarity

Often, proteins may be divided into modular sub-units called domains, which can be associated with specific biological functions. Thus, a protein domain may be considered as the evolutionary unit of biological structure and function [89]. However, while it is well known that the 3D structures of protein domains are often more evolutionarily conserved than their one-dimensional (1D) amino acid sequences, comparing 3D structures is much more difficult than comparing 1D sequences. However, until recently, most evolutionary studies of proteins have compared and clustered 1D amino acid and nucleotide sequences rather than 3D molecular structures.

A pre-requisite for the accurate comparison of protein structures is to have a reliable method for quantifying the structural similarity between pairs of proteins. We recently developed a new protein structure alignment program called Kpax which combines an efficient dynamic programming based scoring function with a simple but novel Gaussian representation of protein backbone shape [76]. This means that we can now quantitatively compare 3D protein domains at a similar rate to throughput to conventional protein sequence comparison algorithms. We recently compared Kpax with a large number of other structure alignment programs, and we found Kpax to be the fastest and amongst the most accurate, in a CATH family recognition test [64]. The latest version of Kpax [9] can calculate multiple flexible alignments, and thus promises to avoid such issues when comparing more distantly related protein folds and fold families.

3.1.3. Formalising and Exploiting Domain Knowledge

Concerning protein structure classification, we aim to explore novel classification paradigms to circumvent the problems encountered with existing hierarchical classifications of protein folds and domains. In particular it will be interesting to set up fuzzy clustering methods taking advantage of our previous work on gene functional classification [50], but instead using Kpax domain-domain similarity matrices. A non-trivial issue with fuzzy clustering is how to handle similarity rather than mathematical distance matrices, and how to find the optimal number of clusters, especially when using a non-Euclidean similarity measure. We will adapt the algorithms and the calculation of quality indices to the Kpax similarity measure. More fundamentally, it will be necessary to integrate this classification step in the more general process leading from data to knowledge called Knowledge Discovery in Databases (KDD) [55].

Another example where domain knowledge can be useful is during result interpretation: several sources of knowledge have to be used to explicitly characterise each cluster and to help decide its validity. Thus, it will be useful to be able to express data models, patterns, and rules in a common formalism using a defined vocabulary for concepts and relationships. Existing approaches such as the Molecular Interaction (MI) format [58] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [69]. A different point of view is represented in the Interaction Network Ontology (INO), a community-driven ontology that aims to standardise and integrate data on interaction networks and to support computer-assisted reasoning [92]. However, this ontology does not integrate basic 3D concepts and structural relationships. Therefore, extending such formalisms and symbolic relationships will be beneficial, if not essential, when classifying the 3D shapes of proteins at the domain family level.

3.1.4. 3D Protein Domain Annotation and Shape Mining

A widely used collection of protein domain families is "Pfam" [54], constructed from multiple alignments of protein sequences. Integrating domain-domain similarity measures with knowledge about domain binding sites, as introduced by us in our KBDOCK approach [2], [4], can help in selecting interesting subsets of domain pairs before clustering. Thanks to our KBDOCK and Kpax projects, we already have a rich set of tools with which we can start to process and compare all known protein structures and PPIs according to their component Pfam domains. Linking this new classification to the latest "SIFTS" (Structure Integration with Function, Taxonomy and Sequence) [86] functional annotations between standard UniProt (http://www.uniprot.org/ sequence identifiers and protein structures from the Protein Data Bank (PDB) [41] could then provide a useful way to discover new structural and functional relationships which are difficult to detect in existing classification schemes such as CATH or SCOP. As part of the thesis project of Seyed Alborzi, we developed a recommender-based data mining technique to associate enzyme classification code numbers with Pfam domains using our recently developed EC-DomainMiner program [1]. We subsequently generalised this approach as a tripartite graph mining method for inferring associations between different protein annotation sources, which we call "CODAC" (for COmputational Discovery of Direct Associations using Common Neighbours). A first paper on CODAC was presented at IWBBIO-2017 [36], and a full paper has recently been accepted by BMC Bioinformatics [13].

3.1.5. Protein Function Annotation

Knowledge of the functional properties of proteins can shed considerable light on how they might interact. However, huge numbers of protein sequences in public databases lack any functional annotation, and the annotation of sequences in such databases is a highly challenging problem. We are developing graph-based and machine learning techniques to annotate automatically the available unannotated sequences in such databases with functional properties such as EC numbers and Gene Ontology (GO) terms. Even if the 3D structures of proteins are unknown, it is natural to suppose that their sequences may be related to each other by the domains, domain families, and super-families that they share. In the frame of the PhD project of Bishnu Sarker, we recently developed a novel graph-based approach called GrAPFI for the automatic functional annotation of protein sequences based on these principles in order to transfer annotations from expert-reviewed sequences to unreviewed sequences in the UniProtKB databases [32], [24].

3.2. Integrative Multi-Component Assembly and Modeling

3.2.1. Context

At the molecular level, each PPI is embodied by a physical 3D protein-protein interface. Therefore, if the 3D structures of a pair of interacting proteins are known, it should in principle be possible for a docking algorithm to use this knowledge to predict the structure of the complex. However, modeling protein flexibility accurately during docking is very computationally expensive due to the very large number of internal degrees of freedom in each protein, associated with twisting motions around covalent bonds. Therefore, it is highly impractical to use detailed force-field or geometric representations in a brute-force docking search. Instead,

most protein docking algorithms use fast heuristic methods to perform an initial rigid-body search in order to locate a relatively small number of candidate binding orientations, and these are then refined using a more expensive interaction potential or force-field model, which might also include flexible refinement using molecular dynamics (MD), for example.

3.2.2. Polar Fourier Docking Correlations

In our *Hex* protein docking program [77], the shape of a protein molecule is represented using polar Fourier series expansions of the form

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

where $\sigma(\underline{x})$ is a 3D shape-density function, a_{nlm} are the expansion coefficients, $R_{nl}(r)$ are orthonormal Gauss-Laguerre polynomials and $y_{lm}(\theta, \phi)$ are the real spherical harmonics. The electrostatic potential, $\phi(\underline{x})$, and charge density, $\rho(\underline{x})$, of a protein may be represented using similar expansions. Such representations allow the *in vacuo* electrostatic interaction energy between two proteins, A and B, to be calculated as [60]

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

This equation demonstrates using the notion of *overlap* between 3D scalar quantities to give a physics-based scoring function. If the aim is to find the configuration that gives the most favourable interaction energy, then it is necessary to perform a six-dimensional search in the space of available rotational and translational degrees of freedom. By re-writing the polar Fourier expansions using complex spherical harmonics, we showed previously that fast Fourier transform (FFT) techniques may be used to accelerate the search in up to five of the six degrees of freedom [78]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [10], [6]. Consequently, we are continuing to explore new ways to exploit the polar Fourier approach.

3.2.3. Assembling Symmetrical Protein Complexes

Although protein-protein docking algorithms are improving [79], [62], it still remains challenging to produce a high resolution 3D model of a protein complex using *ab initio* techniques, mainly due to the problem of structural flexibility described above. However, with the aid of even just one simple constraint on the docking search space, the quality of docking predictions can improve considerably [10], [78]. In particular, many protein complexes involve symmetric arrangements of one or more sub-units, and the presence of symmetry may be exploited to reduce the search space considerably [40], [75], [84]. For example, using our operator notation (in which $\widehat{R}$ and $\widehat{T}$ represent 3D rotation and translation operators, respectively), we have developed an algorithm which can generate and score candidate docking orientations for monomers that assemble into cyclic (C_n) multimers using 3D integrals of the form

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

where the identical monomers A and B are initially placed at the origin, and $\omega_n = 2\pi/n$ is the rotation about the principal n-fold symmetry axis. This example shows that complexes with cyclic symmetry have just 4 rigid body degrees of freedom (DOFs), compared to $6(n-1)$ DOFs for non-symmetrical n-mers. We have generalised these ideas in order to model protein complexes that crystallise into any of the naturally occurring point group symmetries (C_n, D_n, T, O, I). This approach was published in 2016 [8], and was subsequently applied to several symmetrical complexes from the "CAPRI" blind docking experiment [53]. Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to build and refine candidate solutions using a more accurate coarse-grained (CG) force-field scoring function.

3.2.4. Coarse-Grained Models

Many approaches have been proposed in the literature to take into account protein flexibility during docking. The most thorough methods rely on expensive atomistic simulations using MD. However, much of a MD trajectory is unlikely to be relevant to a docking encounter unless it is constrained to explore a putative protein-protein interface. Consequently, MD is normally only used to refine a small number of candidate rigid body docking poses. A much faster, but more approximate method is to use CG normal mode analysis (NMA) techniques to reduce the number of flexible degrees of freedom to just one or a handful of the most significant vibrational modes [68], [52], [65], [66]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [87], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [3].

In the last few years, CG *force-field* models have become increasingly popular in the MD community because they allow very large biomolecular systems to be simulated using conventional MD programs [39]. Typically, a CG force-field representation replaces the atoms in each amino acid with from 2 to 4 "pseudo-atoms", and it assigns each pseudo-atom a small number of parameters to represent its chemo-physical properties. By directly attacking the quadratic nature of pair-wise energy functions, coarse-graining can speed up MD simulations by up to three orders of magnitude. Nonetheless, such CG models can still produce useful models of very large multi-component assemblies [83]. Furthermore, this kind of coarse-graining effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [59]. We are therefore developing a "coarse-grained" scoring function for fast protein-protein docking and multi-component assembly in the frame of the PhD project of Maria-Elisa Ruiz-Echartea [31], [82].

3.2.5. Assembling Multi-Component Complexes and Integrative Structure Modeling

We also want to develop related approaches for integrative structure modeling using cryo-electron microscopy (cryo-EM). Thanks to recently developments in cryo-EM instruments and technologies, it is now feasible to capture low resolution images of very large macromolecular machines. However, while such developments offer the intriguing prospect of being able to trap biological systems in unprecedented levels of detail, there will also come an increasing need to analyse, annotate, and interpret the enormous volumes of data that will soon flow from the latest instruments. In particular, a new challenge that is emerging is how to fit previously solved high resolution protein structures into low resolution cryo-EM density maps. However, the problem here is that large molecular machines will have multiple sub-components, some of which will be unknown, and many of which will fit each part of the map almost equally well. Thus, the general problem of building high resolution 3D models from cryo-EM data is like building a complex 3D jigsaw puzzle in which several pieces may be unknown or missing, and none of which will fit perfectly. We wish to proceed firstly by putting more emphasis on the single-body terms in the scoring function [49], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space (thesis project of Maria Elisa Ruiz Echartea).

DYLISS Project-Team

3. Research Program

3.1. Computer science – symbolic artificial intelligence

We develop methods that use an explicit representation of the relationships between heterogeneous data and knowledge in order to construct a space of hypotheses. Therefore, our objectives in computer science is mainly to develop accurate representations (oriented graphs, Boolean networks, automata, or expressive grammars) to iteratively capture the complexity of a biological system.

Integrating data with querying languages: Semantic web for life sciences The first level of complexity in the data integration process consists in confronting heterogeneous datasets. Both the size and the heretogeneity of life science data make their integration and analysis by domain experts impractical and prone to the streetlight effect (they will pick up the models that best match what they know or what they would like to discover). Our first objective involves the formalization and management of knowledge, that is, the explicitation of relations occurring in structured data. In this setting, our main goal is to facilitate and optimize the integration Semantic Web resources with local users data by relying on the implicit data scheme contained in biological data and Semantic Web resources.

Reasoning over structured data with constraint-based logical paradigms Another level of complexity in life science integration is that very few paradigms exist to model the behavior of a complex biological system. This leads biologists to perform and formulate hypotheses in order to interpret their data. Our strategy is to interpret such hypotheses as combinatorial optimization problems allowing to reduce the family of models compatible with data. To that goal, we collaborate with Potsdam University in order to use and challenge the most recent developments of Answer Set Programming (ASP) [43], a logical paradigm for solving constraint satisfiability and combinatorial optimization issues. Our goal is therefore to provide scalable and expressive formal models of queries on biological networks with the focus of integrating dynamical information as explicit logical constraints in the modeling process.

Characterizing biological sequences with formal syntactic models Our last goal is to identify and characterize the function and expression of genes in non-model species, such as enzymes and isoforms functions in biological networks or specific functional features of metagenomic samples. These are insufficiently precise because of the divergence of biological sequences, the complexity of molecular structures and biological processes, and the weak signals characterizing these elements. Our goal is therefore to develop accurate formal syntactic models (automata, grammars, abstract gene models) enabling us to represent sequence conservation, sets of short and degenerated patterns and crossing or distant dependencies. This requires both to determine classes of formal syntactic models allowing to handle biological complexity, and to automatically characterize the functional potential embodied in biological sequences with these models.

3.2. Scalable methods to query data heterogenity

Confronted to large and complex data sets (raw data are associated with graphs depicting explicit or implicit links and correlations) almost all scientific fields have been impacted by the *big data issue*especially genomics and astronomy [47]. In our opinion, life sciences cumulates several features that are very specific and prevent the direct application of big data strategies that proved successful in other domains such as experimental physics: the existence of **several scales of granularity** from microscopic to macroscopic and the associated issue of dependency propagation, datasets **incompleteness and uncertainty** including highly **heterogeneous** responses to a perturbation from one sample to another, and highly fragmented sources of information that **lacks interoperability** [42]. To explore this research field, we use techniques from symbolic data mining (Semantic Web technologies, symbolic clustering, constraint satisfaction and grammatical modelling) to take into account those life science features in the analysis of biological data.

3.2.1. Research topics

Facilitating data integration and querying The quantity and inner complexity of life science data require semantically-rich analysis methods. A major challenge is then to combine data (from local project as well as from reference databases) and symbolic knowledge seamlessly. Semantic Web technologies provide a relevant framework, as demonstrated by the success of Linked (Open) Data [34]. However, life science end users (1) find it difficult to learn the languages for representing and querying Semantic Web data, and consequently (2) miss the possibility they had to interact with their tabulated data (even when doing so was exceedingly slow and tedious). Our first objective in this axis is to develop accurate abstractions of datasets or knowledge repositories to facilitate their exploration with RDF-based technologies.

Scalability of semantic web queries. A bottleneck in data querying is given by the performance of federated SPARQL queries, which must be improved by several orders of magnitude to allow current massive data to be analyzed. In this direction, our research program focuses the combination of *linked data fragments* [48], query properties and dataset structure for decomposing federated SPARQL queries.

Building and compressing static maps of interacting compounds A final approach to handle heterogeneity is to gather multi-scale data knowledge into functional static map of biological models that can be analyzed and/or compressed. This requires to linking genomics, metabolomics, expression data and protein measurement of several phenotypes into unified frameworks. In this direction, our main goal is to develop families of constraints, inspired by symbolic dynamical systems, to link datasets together. We currently focus on health (personalized medicine) and environmental (role of non-coding regulations, compression) datasets.

3.2.2. Associated software tools

AskOmics platform *AskOmics* is an integration and interrogation software for linked biological data based on semantic web technologies [url]. *AskOmics* aims at bridging the gap between end user data and the Linked (Open) Data cloud (LOD cloud). It allows heterogeneous bioinformatics data (formatted as tabular files or directly in RDF) to be loaded into a Triple Store system using a user-friendly web interface. It helps end users to (1) take advantage of the information readily available in the LOD cloud for analyzing there own data and (2) contribute back to the linked data by representing their data and the associated metadata in the proper format as well as by linking them to other resources. An originality is the graphical interface that allows any dataset to be integrated in a local RDF datawarehouse and SPARQL query to be built transparently and iteratively by a non-expert user.

FinGoc-tools *The FinGoc tools* allow filtering interaction networks with graph-based optimization criteria in order to elucidate the main regulators of an observed phenotype. The main added-value of these tools is to make explicit the criteria used to highlight the role of the main regulators.(1) The KeyRegulatorFinder package searches key regulators of lists of molecules (like metabolites, enzymes or genes) by taking advantage of knowledge databases in cell metabolism and signaling [package]. (2) The PowerGrasp python package implements graph compression methods oriented toward visualization, and based on power graph analysis [package]. (3) The iggy package enables the repairing of an interaction graph with respect to expression data. [Python package]

3.3. Metabolism: from enzyme sequences to systems ecology

Our researches in bioinformatics in relation with metabolic processes are driven by the understanding of non-model (eukaryote) species. Their metabolism have acquired specific features that we wish to identify with computational methods. To that goal, we combine sequence analysis with metabolic network analysis, with the final goal to understand better the metabolism of communities of organisms.

3.3.1. Research topics

Genomic level: characterizing enzymatic functions of protein sequences Precise characterization of functional proteins, such as enzymes or transporters, is a key to better understand and predict the actors involved in a metabolic process. In order to improve the precision of functional annotations, we develop machine learning approaches taking a sample of functional sequences as input to infer a grammar representing their key syntactical characteristics, including dependencies between residues. Our first goal is to enable an automatic semi-supervised refinement of enzymes classification [5] by combining the Protomata-Learner [38] framework - which captures local dependencies - with formal concept analysis. More challenging, we are exploring the learn of grammars representing long-distance dependencies such as those exhibited by contacts of amino-acids that are far in the sequence but close in the 3D protein folding.

System level: enriching and comparing metabolic networks for non-model organisms Non-model organisms are associated with often incomplete and poorly annotated sequences, leading to draft networks of their metabolism which largely suffer from incompleteness. in former studies, the team has developed several methods to improve the quality of eukaryotes metabolic networks, by solving several variants of the so-called *Metabolic Network gap-filling problem* with logical programming approaches [9], [8]. The main drawback of these approaches is that they cannot scale to the reconstruction and comparison of families of metabolic networks. Our main objective is therefore to develop new tools for the comparison of species strains at the metabolic level.

Consortium level: exploring the diversity of community consortia A new emerging field is system ecology, which aims at building predictive models of species interactions within an ecosystem for deciphering cooperative and competitive relationships between species [41].This field raise two new issues (1) uncertainty on the species present in the ecosystem and (2) uncertainty about the global objective governing an ecosystem. To address these challenges, our first research focus is the inference of metabolic exchanges and relationships from transporter identification, based on our expertise in metabolic network gap-filling. A second very challenging focus is the prediction of transporters families by obtaining refined characterization of transporters, which are quite unexplored apart from specific databases [45].

3.3.2. Associated software tools

Protomata is a machine learning suite for the inference of automata characterizing (functional) families of proteins from available sequences by modeling alternative local dependencies. They are well suited to predict new family members with a high specificity [url]. The tool builds sequences alignments (partial and local), learns automata and searches for new family members in sequence databases. Applications of Protomata tools include automatic updating of the cyanolase database [38] and the refinement of the classification of HAD enzymes [5].

AuReMe workspace is designed for tractable reconstruction of metabolic networks [url]. The toolbox allows for the Automatic Reconstruction of Metabolic networks based on the combination of multiple heterogeneous data and knowledge sources [12]. The main added-values are the inclusion of graph-based tools relevant for the study of non-classical organisms (Meneco, Menetools, Shogen packages), the possibility to trace the reconstruction and curation procedures (Padmet package), and the exploration of reconstructed metabolic networks with wikis (wiki-export package). It has been used for reconstructing metabolic networks of micro and macro-algae [44], extremophile bacteria [39] and communities of organisms [3].

3.4. Regulation and signaling: detecting complex and discriminant signatures of phenotypes

On the contrary to metabolic networks, regulatory and signaling processes in biological systems involves agents interacting at different granularity levels (from genes, non-coding RNAs to protein complexes) and different time-scales. Our focus is on the reconstruction of large-scale networks involving multiple scales processes, from which controllers can be extracted with symbolic dynamical systems methods. A particular attention is paid to the characterization of products of genes (such as isoform) and of perturbations to identify discriminant signature of pathologies.

3.4.1. Research topics

Genomic level: characterizing gene structure with grammatical languages and conservation information The subject here is to accurately represent gene structure, including intron/exon structure, for predicting the products of genes, such as isoform transcripts, and comparing the expression potential of a eukaryotic gene according to its context (e.g. tissue) or according to the species. Our approach consists in designing grammatical and comparative-genomics based models for gene structures able to detect heterogeneous functional sites (splicing sites, regulatory binding sites...), functional regions (exons, promotors...) and global constraints (translation into proteins) [35]. Accurate gene models are defined by identifying general constraints shaping gene families and their structures conserved over evolution. Syntactic elements controlling gene expression (transcription factor binding sites controlling transcription; enhancers and silencers controlling splicing events...), that is, short, degenerated and overlapping functional sequences, are modeled by relying on the high capability of SVG grammars to deal with structure and ambiguity [46].

System level: extracting causal signatures of complex phenotypes with systems biology frameworks The main challenge we address is to set up a generic formalism to model inter-layer interactions in large-scale biological networks. To that goal, we have developed several types of abstractions: multi-experiments framework to learn and control signaling networks [10], multi-layer reactions in interaction graphs [36], and multi-layer information in large-scale Petri nets [33]. Our main issues are to scale these approaches to standardized large-scale repositories by relying on the interoperable Linked Open Data (LOD) resources and to enrich them with ad-hoc regulations extracted from sequence-based analysis. This will allow us to characterize changes in system attractors induced by mutations and how they may be included in pathology signatures.

3.4.2. Associated software tools

Logol software is designed for complex pattern modelling and matching [url] It is a swiss-army-knife for pattern matching on DNA/RNA/Protein sequences, based on expressive patterns which consist in a complex combination of motifs (such as degenerated strings) and structures (such as imperfect stem-loop ou repeats) [1]. *Logol* key features are the possibilities (i) to divide a pattern description into several sub-patterns, (ii) to model long range dependencies, and (iii) to enable the use of ambiguous models or to permit the inclusion of negative conditions in a pattern definition. Therefore, *Logol* encompasses most of the features of specialized tools (Vmatch, Patmatch, Cutadapt, HMM) and enables interplays between several classes of patterns (motifs and structures), including stem-loop identification in CRISPR.

Caspo software Cell ASP Optimizer (*Caspo*) constitutes a pipeline for automated reasoning on logical signaling networks (learning, classifying, designing experimental perturbations, identifying controllers, take time-series into account) [url]. The software handles inherent experimental noise but enumerating all different logical networks which are compatible with a set of experimental observations [10]. The main advantage is that is enables a complete study of logical network without requiring any linear constraint programs.

Cadbiom package aims at building and analyzing the asynchronous dynamics of enriched logical networks [url] It is based on Guarded transition semantic and allows synchronization events to be investigated in large-scale biological networks [33]. For instance, it was designed to allow controler of phenotypes in large-scale knowledge databases (PID) to be curated and analyzed [4].

ERABLE Project-Team

3. Research Program

3.1. Two main goals

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

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

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

One important longer term goal of the above is to analyse how the behaviour and dynamics of such a network of networks might be controlled by modifying it, including by subtracting some of its components from the network or by adding new ones.

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

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

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

3.2. Different research axes

The goals of the team are biological and methodological, the two being intrinsically linked. Any division into axes along one or the other aspect or a combination of both is thus somewhat artificial. Following the evaluation of the team at the end of 2017, four main axes were identified, with the last one being the more recently added one. This axis is specifically oriented towards health in general, human or animal. The first three axes are: genomics, metabolism and post-transcriptional regulation, and (c)evolution.

Notice that the division itself is based on the biological level (genomic, metabolic/regulatory, evolutionary) or main current Life Science purpose (health) rather than on the mathematical or computational methodology involved. Any choice has its part of arbitrariness. Through the one we made, we wished to emphasise the fact that the area of application of ERABLE is important for us. *It does not mean that the mathematical and computational objectives are not equally important*, but only that those are, most often, motivated by problems coming from or associated to the general Life Science goal. Notice that such arbitrariness also means that some Life Science topics will be artificially split into two different Axes. One example of this is genomics and the two main health areas currently addressed that are intrinsically inter-related for now.

Axis 1: Genomics
Intra and inter-cellular interactions involve molecular elements whose identification is crucial to understand what governs, and also what might enable to control such interactions. For the sake of clarity, the elements may be classified in two main classes, one corresponding to the elements that allow the interactions to happen by moving around or across the cells, and another that are the genomic regions where contact is established. Examples of the first are non coding RNAs, proteins, and mobile genetic elements such as (DNA) transposons, retro-transposons, insertion sequences, etc. Examples of the second are DNA/RNA/protein binding sites and targets. Furthermore, both types (effectors and targets) are subject to variation across individuals of a population, or even within a single (diploid) individual. Identification of these variations is yet another topic that we wish to cover. Variations are understood in the broad sense and cover single nucleotide polymorphisms (SNPs), copy-number variants (CNVs), repeats other than mobile elements, genomic rearrangements (deletions, duplications, insertions, inversions, translocations) and alternative splicings (ASs). All three classes of identification problems (effectors, targets, variations) may be put under the general umbrella of genomic functional annotation.

Axis 2: Metabolism and post-transcriptional regulation
As increasingly more data about the interaction of molecular elements (among which those described above) becomes available, these should then be modelled in a subsequent step in the form of networks. This raises two main classes of problems. The first is to accurately infer such networks. Assuming such a network, integrated or “simple”, has been inferred for a given organism or set of organisms, the second problem is then to develop the appropriate mathematical models and methods to extract further biological information from such networks.

The team has so far concentrated its efforts on two main aspects concerning such interactions: metabolism and post-transcriptional regulation by small RNAs. The more special niche we have been exploring in relation to metabolism concerns the fact that the latter may be seen as an organism's immediate window into its environment. Finely understanding how species communicate through those windows, or what impact they may have on each other through them is thus important when the ultimate goal is to be able to model communities of organisms, for understanding them and possibly, on a longer term, for control. While such communication has been explored in a number of papers, most do so at a too high level or only considered couples of interacting organisms, not larger communities. The idea of investigating consortia, and in the case of synthetic biology, of using them, has thus started being developed in the last decade only, and was motivated by the fact that such consortia may perform more complicated functions than could single populations, as well as be more robust to environmental fluctuations. Another originality of the work that the team has been doing in the last decade has also been to fully explore the combinatorial aspects of the structures used (graphs or directed hypergraphs) and of the associated algorithms. As concerns post-transcriptional regulation, the team has essentially been exploring the idea that small RNAs may have an important role in the dialog between different species.

Axis 3: (Co)Evolution
Understanding how species that live in a close relationship with others may (co)evolve requires understanding for how long symbiotic relationships are maintained or how they change through time. This may have deep implications in some cases also for understanding how to control such relationships, which may be a way of controlling the impact of symbionts on the host, or the impact of the host on the symbionts and on the environment (by acting on its symbiotic partner(s)). These relationships, also called *symbiotic associations*, have however not yet been very widely studied, at least not at a large scale.

One of the problems is getting the data, meaning the trees for hosts and symbionts but even prior to that, determining with which symbionts the present-day hosts are associated (or are "infected" by as may be the term used in some contexts) which is a big enterprise in itself. The other problem is measuring the stability of the association. This has generally been done by concomitantly studying the phylogenies of hosts and symbionts, that is by doing what is called a *cophylogeny* analysis, which itself is often realised by performing what is called a *reconciliation* of two phylogenetic trees (in theory, it could be more than two but this is a problem that has not yet been addressed by the team), one for the symbionts and one for the hosts with which the symbionts are associated. This consists in mapping one of the trees (usually, the symbiont tree) to the other. Cophylogeny inherits all the difficulties of phylogeny, among which the fact that it is not possible to check the result against the "truth" as this is now lost in the past. Cophylogeny however also brings new problems of its own which are to estimate the frequency of the different types of events that could lead to discrepant evolutionary histories, and to estimate the duration of the associations such events may create.

Axis 4: Human, animal and plant health
As indicated above, this is a recent axis in the team and concerns various applications to human and animal health. In some ways, it overlaps with the three previous axes as well as with Axis 5 on the methodological aspects, but since it gained more importance in the past few years, we decided to develop more these particular applications. Most of them started through collaborations with clinicians. Such applications are currently focused on three different topics: (i) Infectiology, (ii) Rare diseases, and (iii) Cancer.

Infectiology is the oldest one. It started by a collaboration with Arnaldo Zaha from the Federal University of Rio Grande do Sul in Brazil that focused on pathogenic bacteria living inside the respiratory tract of swines. Since our participation in the H2020 ITN MicroWine, we started interested in infections affecting plants this time, and more particularly vine plants. Rare Diseases on the other hand started by a collaboration with clinicians from the Centre de Recherche en Neurosciences of Lyon (CNRL) and is focused the Taybi-Linder Syndrome (TALS) and on abnormal splicing of U12 introns, while Cancer rests on a collaboration with the Centre Léon Bérard (CLB) and Centre de Recherche en Cancérologie of Lyon (CRCL) which is focused on Breast and Prostate carcinomas and Gynaecological carcinosarcomas.

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

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

GENSCALE Project-Team

3. Research Program

3.1. Axis 1: Data Structure

The aim of this axis is to develop efficient data structures for representing the mass of genomic data generated by the sequencing machines. This research is motivated by the fact that the treatments of large genomes, such as mammalian or plant genomes, require high computing resources, and more specifically very important memory configuration. For example, the ABYSS software used 4.3TB of memory to assemble the white spruce genome [36]. The main reason for such memory consumption is that the data structures used in ABYSS are far from optimal (and this is also the case for many assembly software).

Our research focuses on the de-Bruijn graph structure. This well-known data structure, directly built from raw sequencing data, have many properties matching perfectly well with NGS (Next Generation Sequencing) processing requirements (see next section). Here, the question we are interested in is how to provide a low memory footprint implementation of the de-Bruijn graph to process very large NGS datasets, including metagenomic ones [3], [4].

Another research direction of this axis is the indexing of large sets of objects. A typical, but non exclusive, need is to annotate nodes of the de-Bruijn graph, that is potentially billions of items. Again, very low memory footprint indexing structures are mandatory to manage a very large quantity of objects [5].

3.2. Axis 2: Algorithms

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

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

- **NGS data Compression** De-Bruijn graphs are de facto a compressed representation of the NGS information from which very efficient and specific compressors can be designed. Furthermore, compressing the data using smart structures may speed up some downstream graph-based analyses since a graph structure is already built [1].
- **Genome assembly** This task remains very complicated, especially for large and complex genomes, such as plant genomes with polyploid and highly repeated structures. We worked both on the generation of contigs [3] and on the scaffolding step [26].
- **Detection of variants** This is often the first information we want to extract from billions of reads. Variant structures range from SNPs or short indels to large insertions/deletions and long inversions over the chromosomes. We developed original methods to find variants without any reference genome [7]. We also worked on the detection of structural variants using approaches of local assembly [6].
- **Metagenomics** We focussed our research on comparative metagenomics by providing methods able to compare hundreds of metagenomic samples together. This is achieved by combining very low memory data structures and efficient implementation and parallelization on large clusters [2].
- **Genome Wide Association Study (GWAS)** We tackle this problem with algorithms commonly used in data mining. From two cohorts of individuals (case and control) we can exhibit statistically significant *patterns* spaning over full genomes.

In addition, we also proposed new algorithmic solutions for analyzing third generation sequencing data, in order to benefit from their larger read size while taking into account their higher sequencing error rate [16].

3.3. Axis 3: Parallelism

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

GenScale does not do explicit research on parallelism. It exploits the capacity of computing resources to support parallelism. The problem is addressed in two different directions. The first is an engineering approach that uses existing parallel tools to implement algorithms such as multithreading or MapReduce techniques [4]. The second is a parallel algorithmic approach: during the development step, the algorithms are constrained by parallel criteria [2]. This is particularly true for parallel algorithms targeting hardware accelerators.

IBIS Project-Team

3. Research Program

3.1. Analysis of qualitative dynamics of gene regulatory networks

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

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

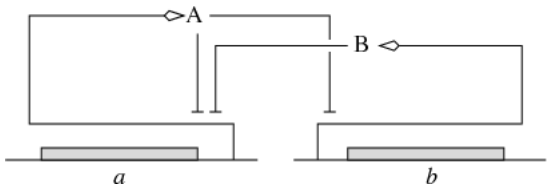
A variety of qualitative modeling formalisms have been introduced over the past decades. Boolean or logical models, which describe gene regulatory and signalling networks as discrete-time finite-state transition systems, are probably most widely used. The dynamics of these systems are governed by logical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

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

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

Participants: Eugenio Cinquemani [Correspondent], Johannes Geiselmann, Hidde de Jong, Stephan Lacour, Aline Marguet, Michel Page, Corinne Pinel, Delphine Ropers.

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



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

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

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

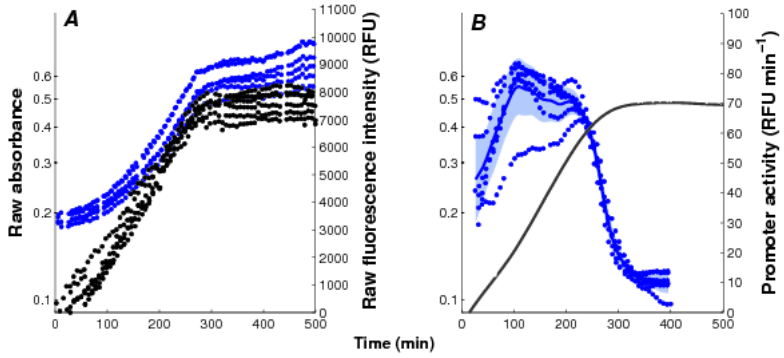


Figure 3. Monitoring of bacterial gene expression in vivo using fluorescent reporter genes (Stefan et al., PLoS Computational Biology, 11(1):e1004028, 2015). The plots show the primary data obtained in a kinetic experiment with E. coli cells, focusing on the expression of the motility gene tar in a mutant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

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

3.3. Analysis of integrated metabolic and gene regulatory networks

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

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

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

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

3.4. Natural and engineered control of growth and gene expression

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

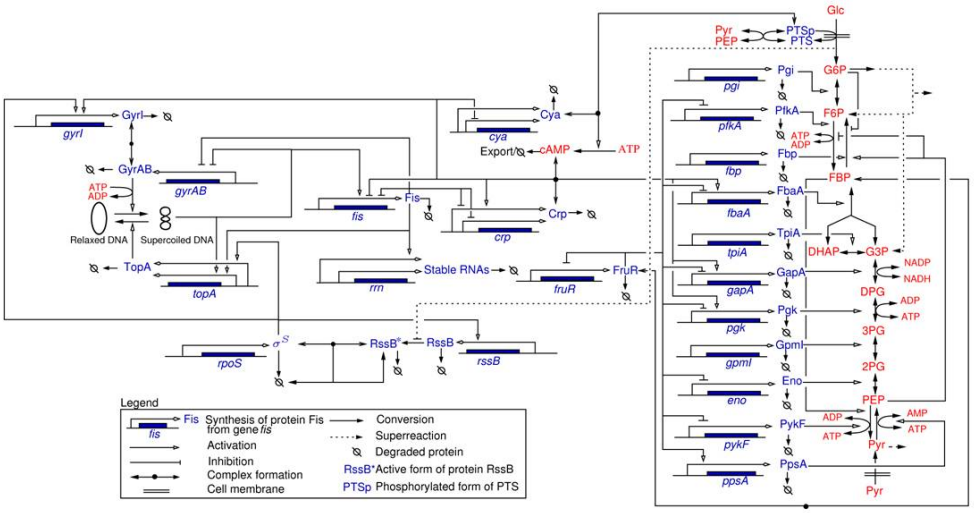


Figure 4. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in E. coli (Baldazzi et al., PLoS Computational Biology, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the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

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

A complementary perspective consists in the use of control-theoretical approaches to modify the functioning of a bacterial cell towards a user-defined objective, by rewiring and selectively perturbing its regulatory networks. The question how regulatory networks in microorganisms can be externally controlled using engineering approaches has a long history in biotechnology and is receiving much attention in the emerging field of synthetic biology. Within a number of on-going projects, IBIS is focusing on two different questions. The first concerns the development of open-loop and closed-loop growth-rate controllers of bacterial cells for both fundamental research and biotechnological applications (Figure 5). Second, we are working on the development of methods for the real-time control of the expression of heterologous proteins in communities of interacting bacterial populations. The above projects involve collaborations with, among others, the Inria project-teams LIFEWARE (INBIO), BIOCORE, and McTAO as well as with a biophysics group at Univ Paris Descartes and a mathematical modeling group at INRA Jouy-en-Josas.

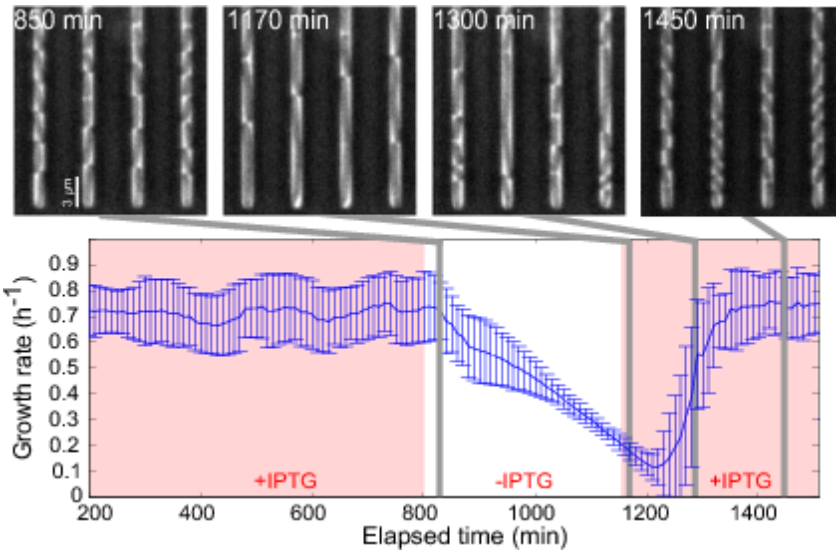


Figure 5. Growth arrest by external control of the gene expression machinery (Izard, Gomez Balderas et al., Molecular Systems Biology, 11:840, 2015). An E. coli strain in which an essential component of the gene expression machinery, the $\beta\beta'$ subunits of RNA polymerase, was put under the control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

LIFEWARE Project-Team

3. Research Program

3.1. Computational Systems Biology

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

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

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

3.2. Chemical Reaction Network (CRN) Theory

Feinberg's chemical reaction network theory and Thomas's influence network analyses provide sufficient and/or necessary structural conditions for the existence of multiple steady states and oscillations in regulatory networks. Those conditions can be verified by static analyzers without knowing kinetic parameter values nor making any simulation. In this domain, most of our work consists in analyzing the interplay between the **structure** (Petri net properties, influence graph, subgraph epimorphisms) and the **dynamics** (Boolean, CTMC, ODE, time scale separations) of biochemical reaction systems. In particular, our study of influence graphs of reaction systems, our generalization of Thomas' conditions of multi-stationarity and Soulé's proof to reaction systems [0], the inference of reaction systems from ODEs [0], the computation of structural invariants by constraint programming techniques, and the analysis of model reductions by subgraph epimorphisms now provide solid ground for developing static analyzers, using them on a large scale in systems biology, and elucidating modules.

3.3. Logical Paradigm for Systems Biology

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

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

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

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

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

In particular, the definition of a continuous satisfaction degree for **first-order temporal logic** formulae with constraints over the reals, was the key to generalize this approach to quantitative models, opening up the field of model-checking to model optimization [0] This line of research continues with the development of temporal logic patterns with efficient constraint solvers and their generalization to handle stochastic effects.

3.4. Computer-Aided Design of CRNs for Synthetic Biology

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

3.5. Modeling of Phenotypic Heterogeneity in Cellular Processes

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

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

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

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

[0] Fages, François, Le Guludec, Guillaume and Bournez, Olivier, Pouly, Amaury. Strong Turing Completeness of Continuous Chemical Reaction Networks and Compilation of Mixed Analog-Digital Programs. In CMSB’17: Proceedings of the fiveteen international conference on Computational Methods in Systems Biology, pages 108–127, volume 10545 of Lecture Notes in Computer Science. Springer-Verlag, 2017.

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

Appropriate mathematical and computational tools should then be employed for the analysis of these models and their calibration to experimental data. One can notably mention **global optimization** tools to search for appropriate parameters within large spaces, **moment closure** approaches to efficiently approximate stochastic models [0], and (stochastic approximations of) the **expectation maximization** algorithm for the identification of mixed-effects models [0].

3.6. External Control of Cell Processes

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

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

More recently, motivated by an analogy with a benchmark control problem, the stabilization of an inverted pendulum, we investigated the possibility to balance a genetic toggle switch in the vicinity of its unstable equilibrium configuration. We searched for solutions to balance an individual cell and even an entire population of heterogeneous cells, each harboring a toggle switch [0].

Independently, in collaboration with colleagues from IST Austria, we investigated the problem of controlling cells, one at a time, by constructing an integrated optogenetic-enabled microscopy platform. It enables experiments that bridge individual and population behaviors. We demonstrated: (i) population structuring by independent closed-loop control of gene expression in many individual cells, (ii) cell–cell variation control during antibiotic perturbation, (iii) hybrid bio-digital circuits in single cells, and freely specifiable digital communication between individual bacteria [0].

3.7. Constraint Solving and Optimization

Constraint solving and optimization methods are important in our research. On the one hand, static analysis of biochemical reaction networks involves solving hard combinatorial optimization problems, for which **constraint programming** techniques have shown particularly successful, often beating dedicated algorithms

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

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

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

[0]Jean-Baptiste Lugagne, Sebastian Sosa Carrillo and Melanie Kirch, Agnes Köhler, Gregory Batt and Pascal Hersen. Balancing a genetic toggle switch by real-time feedback control and periodic forcing. Nature Communications, 8(1):1671, 2017.

[0]Remy Chait, Jakob Ruess, Tobias Bergmiller and Gavsper Tkavcik, Cvalin Guet. Shaping bacterial population behavior through computer-interfaced control of individual cells. Nature Communications, 8(1):1535, 2017.

and allowing to solve large instances from model repositories. On the other hand, parameter search and model calibration problems involve similarly solving hard continuous optimization problems, for which **evolutionary algorithms**, and especially the covariance matrix evolution strategy (CMA-ES) [0] have been shown to provide best results in our context, for up to 100 parameters. This has been instrumental in building challenging quantitative models, gaining model-based insights, revisiting admitted assumptions, and contributing to biological knowledge [00].

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

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

[0] Pauline Traynard, Céline Feillet, Sylvain Soliman, Franck Delaunay, François Fages. Model-based Investigation of the Circadian Clock and Cell Cycle Coupling in Mouse Embryonic Fibroblasts: Prediction of RevErb-alpha Up-Regulation during Mitosis. Biosystems, 149:59–69, 2016.

MORPHEME Project-Team

3. Research Program

3.1. Research program

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

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

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

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

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

MOSAIC Team

3. Research Program

3.1. Axis1: Representation of biological organisms and their forms in silico

The modeling of organism development requires a formalization of the concept of form, *i.e.* a mathematical definition of what is a form and how it can change in time, together with the development of efficient algorithms to construct corresponding computational representations from observations, to manipulate them and associate local molecular and physical information with them. Our aim is threefold. First, we will develop new computational structures that make it possible to represent complex forms efficiently in space and time. For branching forms, the challenge will be to reduce the computational burden of the current tree-like representations that usually stems from their exponential increase in size during growth. For tissue structures, we will seek to develop models that integrate seamlessly continuous representations of the cell geometry and discrete representations of their adjacency network in dynamical and adaptive framework. Second, we will explore the use of machine learning strategies to set up robust and adaptive strategies to construct form representations in computers from imaging protocols. Finally, we will develop the notion of digital atlases of development, by mapping patterns of molecular (gene activity, hormones concentrations, cell polarity, ...) and physical (stress, mechanical properties, turgidity, ...) expressions observed at different stages of development on models representing average form development and by providing tools to manipulate and explore these digital atlases.

3.2. Axis2: Data-driven models of form development

Our aim in this second research axis will be to develop models of physiological patterning and bio-physical growth to simulate the development of 3D biological forms in a realistic way. Models of key processes participating to different aspects of morphogenesis (signaling, transport, molecular regulation, cell division, etc.) will be developed and tested *in silico* on 3D data structures reconstructed from digitized forms. The way these component-based models scale-up at more abstract levels where forms can be considered as continuums will also be investigated. Altogether, this will lead us to design first highly integrated models of form development, combining models of different processes in one computational structure representing the form, and to analyze how these processes interact in the course of development to build up the form. The simulation results will be assessed by quantitative comparison with actual form development. From a computational point of view, as branching or organ forms are often represented by large and complex data-structures, we aim to develop optimized data structures and algorithms to achieve satisfactory compromises between accuracy and efficiency.

3.3. Axis3: Plasticity and robustness of forms

In this research axis, building on the insights gained from axes 1 and 2 on the mechanisms driving form development, we aim to explore the mechanistic origin of form plasticity and robustness. At the ontogenetic scale, we will study the ability of specific developmental mechanisms to buffer, or even to exploit, biological noise during morphogenesis. For plants, we will develop models capturing morphogenetic reactions to specific environmental changes (such as water stress or pruning), and their ability to modulate or even to reallocate growth in an opportunistic manner.

At the phylogenetic scale, we will investigate new connections that can be drawn from the use of a better understanding of form development mechanisms in the evolution of forms. In animals, we will use ascidians as a model organism to investigate how the variability of certain genomes relates to the variability of their forms. In plants, models of the genetic regulation of form development will be used to test hypotheses on the evolution of regulatory gene networks of key morphogenetic mechanisms such as branching. We believe that a better mechanistic understanding of developmental processes should shed new light on old evo-devo questions related to the evolution of biological forms, such as understanding the origin of *developmental constraints*[0] how the internal rules that govern form development, such as chemical interactions and physical constraints, may channel form changes so that selection is limited in the phenotype it can achieve?

3.4. Key modeling challenges

During the project lifetime, we will address several computational challenges related to the modeling of living forms and transversal to our main research axes. During the first phase of the project, we concentrate on 4 key challenges.

3.4.1. A new paradigm for modeling tree structures in biology

There is an ubiquitous presence of tree data in biology: plant structures, tree-like organs in animals (lungs, kidney vasculature), corals, sponges, but also phylogenic trees, cell lineage trees, *etc*. To represent, analyze and simulate these data, a huge variety of algorithms have been developed. For a majority, their computational time and space complexity is proportional to the size of the trees. In dealing with massive amounts of data, like trees in a plant orchard or cell lineages in tissues containing several thousands of cells, this level of complexity is often intractable. Here, our idea is to make use of a new class of tree structures, that can be efficiently compressed and that can be used to approximate any tree, to cut-down the complexity of usual algorithms on trees.

3.4.2. Efficient computational mechanical models of growing tissues

The ability to simulate efficiently physical forces that drive form development and their consequences in biological tissues is a critical issue of the MOSAIC project. Our aim is thus to design efficient algorithms to compute mechanical stresses within data-structures representing forms as the growth simulation proceeds. The challenge consists of computing the distribution of stresses and corresponding tissue deformations throughout data-structures containing thousands of 3D cells in close to interactive time. For this we will develop new strategies to simulate mechanics based on approaches originally developed in computer graphics to simulate in real time the deformation of natural objects. In particular, we will study how meshless and isogeometric variational methods can be adapted to the simulation of a population of growing and dividing cells.

3.4.3. Realistic integrated digital models

Most of the models developed in MOSAIC correspond to specific parts of real morphogenetic systems, avoiding the overwhelming complexity of real systems. However, as these models will be developed on computational structures representing the detailed geometry of an organ or an organism, it will be possible to assemble several of these sub-models within one single model, to figure out missing components, and to test potential interactions between the model sub-components as the form develops.

Throughout the project, we will thus develop two digital models, one plant and one animal, aimed at integrating various aspects of form development in a single simulation system. The development of these digital models will be made using an agile development strategy, in which the models are created and get functional at a very early stage, and become subsequently refined progressively.

[0] Raff, R. A. (1996). The Shape of Life: Genes, Development, and the Evolution of Form. Univ. Chicago Press.

3.4.4. Development of a computational environment for the simulation of biological form development

To support and integrate the software components of the team, we aim to develop a computational environment dedicated to the interactive simulation of biological form development. This environment will be built to support the paradigm of dynamical systems with dynamical structures. In brief, the form is represented at any time by a central data-structure that contains any topological, geometric, genetic and physiological information. The computational environment will provide in a user-friendly manner tools to up-load forms, to create them, to program their development, to analyze, visualize them and interact with them in 3D+time.

PLEIADE Team

3. Research Program

3.1. A Geometric View of Diversity

Diversity may be studied as a set of dissimilarities between objects. The underlying mathematical construction is the notion of distance. Knowing a set of objects, it is possible, after computation of pairwise distances, or sometimes dissimilarities, to build a Euclidean image of it as a point cloud in a space of relevant dimension. Then, diversity can be associated with the shape of the point cloud. The human eye is often far better than an algorithm at recognizing a pattern or shape. One objective of our project is to narrow the gap between the story that a human eye can tell, and that an algorithm can tell. Several directions will be explored. First, this requires mastering classical tools in dimension reduction, mainly algebraic tools (PCA, NGS, Isomap, eigenmaps, etc ...). Second, neighborhoods in point clouds naturally lead to graphs describing the neighborhood networks. There is a natural link between modular structures in distance arrays and communities on graphs. Third, points (representing, say, DNA sequences) are samples of diversity. Dimension reduction may show that they live on a given manifold. This leads to geometry (differential or Riemannian geometry). It is expected that some properties of the manifold can tell something of the constraints on the space where measured individuals live. The connection between Riemannian geometry and graphs, where weighted graphs are seen as mesh embedded in a manifold, is currently an active field of research [28], [27]. See as well [30] for a link between geometric structure, linear and nonlinear dimensionality reduction.

Biodiversity and high-performance computing: Most methods and tools for characterizing diversity have been designed for datasets that can be analyzed on a laptop, but NGS datasets produced for metabarcoding are far too large. Data analysis algorithms and tools must be revisited and scaled up. We will mobilize both distributed algorithms like the Arnoldi method and new algorithms, like random projection or column selection methods, to build point clouds in Euclidean spaces from massive data sets, and thus to overcome the cubic complexity of computation of eigenvectors and eigenvalues of very large dense matrices. We will also link distance geometry [22] with convex optimization procedures through matrix completion [16], [17].

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

3.2. Knowledge Management for Biology

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

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

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

3.3. Modeling by successive refinement

Describing the links between diversity in traits and diversity in function will require comprehensive models, assembled from and refining existing models. A recurring difficulty in building comprehensive models of biological systems is that accurate models for subsystems are built using different formalisms and simulation techniques, and hand-tuned models tend to be so focused in scope that it is difficult to repurpose them [14]. Our belief is that a sustainable effort in building efficient behavioral models must proceed incrementally, rather than by modeling individual processes *de novo*. *Hierarchical modeling* [11] is one way of combining specific models into networks. Effective use of hierarchical models requires both formal definition of the semantics of such composition, and efficient simulation tools for exploring the large space of complex behaviors. We have previously shown that this approach can be effective for certains kinds of systems in biotechnology [2], [15] and medicine [13]. Our challenge is to adapt incremental, hierarchical refinement to modeling organisms and communities in metagenomic and comparative genomic applications.

SERPICO Project-Team

3. Research Program

3.1. Statistics and algorithms for computational microscopy

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

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

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

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

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

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

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

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

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

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

ARAMIS Project-Team

3. Research Program

3.1. From geometrical data to multimodal imaging

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

3.2. Models of brain networks

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

3.3. Spatiotemporal modeling from longitudinal data

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

3.4. Decision support systems

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

3.5. Clinical research studies

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

ATHENA Project-Team

3. Research Program

3.1. Computational diffusion MRI

Diffusion MRI (dMRI) provides a non-invasive way of estimating in-vivo CNS fiber structures using the average random thermal movement (diffusion) of water molecules as a probe. It's a recent field of research with a history of roughly three decades. It was introduced in the mid 80's by Le Bihan et al [71], Merboldt et al [77] and Taylor et al [88]. As of today, it is the unique non-invasive technique capable of describing the neural connectivity in vivo by quantifying the anisotropic diffusion of water molecules in biological tissues.

3.1.1. Diffusion Tensor Imaging & High Angular Resolution Diffusion Imaging

In dMRI, the acquisition and reconstruction of the diffusion signal allows for the reconstruction of the water molecules displacement probability, known as the Ensemble Average Propagator (EAP) [87], [53]. Historically, the first model in dMRI is the 2nd order diffusion tensor (DTI) [51], [50] which assumes the EAP to be Gaussian centered at the origin. DTI (Diffusion Tensor Imaging) has now proved to be extremely useful to study the normal and pathological human brain [72], [62]. It has led to many applications in clinical diagnosis of neurological diseases and disorder, neurosciences applications in assessing connectivity of different brain regions, and more recently, therapeutic applications, primarily in neurosurgical planning. An important and very successful application of diffusion MRI has been brain ischemia, following the discovery that water diffusion drops immediately after the onset of an ischemic event, when brain cells undergo swelling through cytotoxic edema.

The increasing clinical importance of diffusion imaging has driven our interest to develop new processing tools for Diffusion Tensor MRI. Because of the complexity of the data, this imaging modality raises a large amount of mathematical and computational challenges. We have therefore developed original and efficient algorithms relying on Riemannian geometry, differential geometry, partial differential equations and front propagation techniques to correctly and efficiently estimate, regularize, segment and process Diffusion Tensor MRI (DT-MRI) (see [74] and [73]).

In DTI, the Gaussian assumption over-simplifies the diffusion of water molecules. While it is adequate for voxels in which there is only a single fiber orientation (or none), it breaks for voxels in which there are more complex internal structures and limitates the ability of the DTI to describe complex, singular and intricate fiber configurations (U-shape, kissing or crossing fibers). To overcome this limitation, so-called Diffusion Spectrum Imaging (DSI) [91] and High Angular Resolution Diffusion Imaging (HARDI) methods such as Q-ball imaging [89] and other multi-tensors and compartment models [84], [86], [44], [43], [80] were developed to resolve the orientationnality of more complicated fiber bundle configurations.

Q-Ball imaging (QBI) has been proven very successful in resolving multiple intravoxel fiber orientations in MR images, thanks to its ability to reconstruct the Orientation Distribution Function (ODF, the probability of diffusion in a given direction). These tools play a central role in our work related to the development of a robust and linear spherical harmonic estimation of the HARDI signal and to our development of a regularized, fast and robust analytical QBI solution that outperforms the state-of-the-art ODF numerical technique developed by Tuch [89]. Those contributions are fundamental and have already started to impact on the Diffusion MRI, HARDI and Q-Ball Imaging community [61]. They are at the core of our probabilistic and deterministic tractography algorithms devised to best exploit the full distribution of the fiber ODF (see [58], [3] and [59], [4]).

3.1.2. Beyond DTI with high order tensors

High Order Tensors (HOT) models to estimate the diffusion function while overcoming the shortcomings of the 2nd order tensor model have also been recently proposed such as the Generalized Diffusion Tensor Imaging (G-DTI) model developed by Ozarslan et al [95], [96] or 4th order Tensor Model [49]. For more details, we refer the reader to our articles in [64], [84] where we review HOT models and to our articles in [73], co-authored with some of our close collaborators, where we review recent mathematical models and computational methods for the processing of Diffusion Magnetic Resonance Images, including state-of-the-art reconstruction of diffusion models, cerebral white matter connectivity analysis, and segmentation techniques. Recently, we started to work on Diffusion Kurtosis Imaging (DKI), of great interest for the company OLEA MEDICAL(http://www.olea-medical.com/). Indeed, DKI is fastly gaining popularity in the domain for characterizing the diffusion propagator or EAP by its deviation from Gaussianity. Hence it is an important clinical tool for characterizing the white-matter's integrity with biomarkers derived from the 3D 4th order kurtosis tensor (KT) [67].

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 [43], [44]. 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 increases, 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 [66], [84].

3.1.3. Improving dMRI acquisitions

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

First, we contributed developing real-time reconstruction algorithm based on the Kalman filter [57]. 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 [78].

We have contributed to the reconstruction of the diffusion signal and its important features as the orientation distribution function and the ensemble average propagator, with a special focus on clinical setting in particular for single and multiple Q-shell experiments. 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 [78], [54], [56].

We have also contributed to design optimal acquisition schemes for single and multiple Q-shell experiments. In particular, the method proposed in [2] helps generate sampling schemes with optimal angular coverage for multi-shell acquisitions. The cost function we proposed is an extension of the electrostatic repulsion to multi-shell and can be used to create acquisition schemes with incremental angular distribution, compatible with prematurely stopped scans. Compared to more commonly used radial sampling, our method improves the

angular resolution, as well as fiber crossing discrimination. The optimal sampling schemes, freely available for download[0], have been selected for use in the HCP (Human Connectome Project)[0].

We think that such kind of contributions open new perspectives for dMRI applications including, for example, tractography where the improved characterization of the fiber orientations is likely to greatly and quickly help tracking through regions with and/or without crossing fibers [65].

3.1.4. dMRI modelling, tissue microstructures features recovery & applications

The dMRI signal is highly complex, hence, the mathematical tools required for processing it have to be commensurate in their complexity. Overall, these last twenty years have seen an explosion of intensive scientific research which has vastly improved and literally changed the face of dMRI. In terms of dMRI models, two trends are clearly visible today: the parametric approaches which attempt to build models of the tissue to explain the signal based on model-parameters such as CHARMED [45], AxCaliber [46] and NODDI [92] to cite but a few, and the non-parametric approaches, which attempt to describe the signal in useful but generic functional bases such as the Spherical Polar Fourier (SPF) basis [48], [47], the Solid Harmonic (SoH) basis [60], the Simple Harmonic Oscillator based Reconstruction and Estimation (SHORE) basis [93] and more recent Mean Apparent Propagator or MAP-MRI basis [94].

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

3.1.5. Towards microstructural based tractography

In order to go far beyond traditional fiber-tracking techniques, we believe that first order information, i.e. fiber orientations, has to be superseeded by second and third order information, such as microstructure details, to improve tractography. However, many of these higher order information methods are relatively new or unexplored and tractography algorithms based on these high order based methods have to be conceived and designed. In this aim, we propose to work with multiple-shells to reconstruct the Ensemble Average Propagator (EAP), which represents the whole 3D diffusion process and use the possibility it offers to deduce valuable insights on the microstructural properties of the white matter. Indeed, from a reconstructed EAP one can compute the angular features of the diffusion in an diffusion Orientation Distribution Function (ODF), providing insight in axon orientation, calculate properties of the entire diffusion in a voxel such as the Mean Squared Diffusivity (MSD) and Return-To-Origin Probability (RTOP), or come forth with bio-markers detailing diffusion along a particular white matter bundle direction such as the Return-to-Axis or Return-to-Plane Probability (RTAP or RTPP). This opens the way to a ground-breaking computational and unified framework for tractography based on EAP and microstructure features [6]. Using additional a priori anatomical and/or functional information, we could also constrain the tractography algorithm to start and terminate the streamlines only at valid processing areas of the brain.

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

[0] http://www.emmanuelcaruyer.com/
[0] http://humanconnectome.org/documentation/Q1/imaging-protocols.html

3.1.6. Going beyond the state-of-the-art dMRI

Overall, these last twenty years have seen an explosion of intensive scientific research which has vastly improved and literally changed the face of dMRI.

However, although great improvements have been made, major improvements are still required primarily to optimally acquire dMRI data, better understand the biophysics of the signal formation, recover invariant and intrinsic microstructure features, identify bio-physically important bio-markers and improve tractography. For short, there

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

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 (Superconducting QUantum Interference Device) 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 (Functional MRI) and SPECT (Single-Photon Emission Computed Tomography) 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 [70], [7] and means to calibrate them [90] 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.

3.3. Combined M/EEG and dMRI

dMRI provides a global and systematic view of the long-range structural connectivity within the whole brain. In particular, it allows the recovery of the fiber structure of the white matter which can be considered as the wiring connections between distant cortical areas. These white matter based tractograms are analyzed e.g. to explore the differences in structural connectivity between pathological and normal populations. Moreover, as a by-product, the tractograms can be processed to reveal the nodes of the brain networks, i.e. by segregating together gray matter that share similar connections to the rest of the white matter. But dMRI does not provide information on:

- the cortico-cortical pathways (not passing through white matter) and to some extent, on the short-range connections in the white matter,
- the actual use of connections over time during a given brain activity.

On the opposite, M/EEG measures brain activation over time and provides, after source reconstruction (solving the so-called inverse problem of source reconstruction), time courses of the activity of the cortical areas. Unfortunately, deep brain structures have very little contribution to M/EEG measurements and are thus difficult to analyze. Consequently, M/EEG reveals information about the nodes of the network, but in a more blurry (because of the inverse problem) and fragmented view than dMRI (since it can only reveal brain areas measurable in M/EEG whose activity varies during the experimental protocol). Given its very high temporal resolution, the signal of reconstructed sources can be processed to reveal the functional connectivity between the nodes [85].

While dMRI and M/EEG have been the object of considerable research separately, there have been very few studies on combining the information they provide. Some existing studies deal with the localization of abnormal MEG signals, particularly in the case of epilepsy, and on studying the white matter fibers near the detected abnormal source [76], [79], but to our knowledge there are very few studies merging data coming both from M/EEG and dMRI at the analysis level [82], [63], [52], [83].

Combining the structural and functional information provided by dMRI and M/EEG is a difficult problem as the spatial and temporal resolutions of the two types of measures are extremely different. Still, combining the measurements obtained by these two types of techniques has the great potential of providing a detailed view both in space and time of the functioning brain at a macroscopic level. Consequently, it is a timely and extremely important objective to develop innovative computational tools and models that advance the dMRI and M/EEG state-of-the-art and combine these imaging modalities to build a comprehensive dynamical structural-functional brain connectivity network to be exploited in brain connectivities diseases.

The COBCOM ERC project aims to develop a joint dynamical structural-functional brain connectivity network built on advanced and integrated dMRI and M/EEG ground-breaking methods. To this end, COBCOM will provide new generation of computational dMRI and M/EEG models and methods for identifying and characterizing the connectivities on which the joint network is built. Capitalizing on the strengths of dMRI & M/EEG and building on the bio-physical and mathematical foundations of our models, COBCOM will contribute to create a joint and solid network which will be exploited to identify and characterize white matter abnormalities in some high-impact brain diseases such as Multiple Sclerosis (MS), Epilepsy and mild Traumatic Brain Injury (mTBI).

BIOVISION Project-Team

3. Research Program

3.1. Introduction

The Biovision team has started on January 1st, 2016 and became an Equipe Projet Inria on August 1st, 2018 . It aims at developing fundamental research as well as technological developments along two axes.

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

Visual impairment, also known as vision loss, is a decreased ability to see to a degree that causes problems not fixable by usual means, such as glasses or lenses. Low-vision is a condition caused by eye disease, in which visual acuity is 20/70, meaning that the person is able to see, at 20 meters from a chart, what a normal person would see at 70 meters. Visual impairment affects some 285 million humans in the world, mostly in developed countries where this number is going to increase rapidly due to aging. 85% have low-vision or poorer. [0] There is a strong need to conceive new aid-systems to help these people in their daily living activities. Such systems already exist and can be divided into two categories according to their function. The first category concerns aids that translate visual information into alternative sensory information, such as touch or sound, called Sensory Substitution Devices (SSDs) [34], [31]. The second category concerns aids that adapt visual information to render it more visible to the patients, using scene processing methods and suitable devices. These are based on technological and algorithmic solutions that enhance salient scene characteristics [53], [43]. In Biovision team, we focus on this second category by targeting new vision aid-systems helping patients in their daily life, adapting to their own pathology.

We have strong contacts and collaborations with low-vision centers and associations in order to better understand low-vision patients needs, and have feedback on our prototypes aimed to be distributed to patients via transfer or company creation (startup). Our goal is to develop solutions based on head mounted displays, especially low cost and large public systems, with full consideration of comfort and **ergonomics**. In particular, we focus on three main goals:

1. Developing **reading aids in virtual reality**. This includes functional vision testing, allowing display and interaction to be personalized.
2. Developing broader **vision aid-systems in augmented reality** for other daily living activities, such as social interaction, visual search and navigation.
3. Proposing **image enhancements** which can be customized for each patient depending on their needs and **pathology**.

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

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

[0] Source: VisionAware

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

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

Finally, although this is not the main goal of our team, another natural avenue of our research is to develop novel synergistic solutions to solve computer vision tasks based on bio-inspired mechanisms [7].

3.2. Scientific methodology

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

3.2.1. Adaptive image enhancement

Image enhancement is a natural type of image processing method to help low-vision people better understand visual scenes. An impressive number of techniques have been developed in the fields of computer vision and computer graphics to manipulate image content for a variety of applications. Some of these methods have a direct interest in the design of vision aid-systems. Only a few of them have been carefully evaluated with patients [28], [38], [39], [32], [29]. Our objective is to further exploit and evaluate them with patients, considering dedicated use-cases, using virtual and augmented reality technology (Sec. 3.2.2). We consider not only classical brightness manipulations (e.g., equalization, gamma correction, tone mapping, edge enhancement, image decomposition and cartoonization) but also more sophisticated approaches which can change the geometric information of the scene to highlight the most relevant informations (e.g., scene retargeting and seam carving). In addition, we investigate how image enhancements could be adapted to patients needs by relating tuning parameters to the patient pathology.

3.2.2. Virtual, mixed and augmented reality

Virtual, mixed and augmented reality technology (VR/MR/AR) is based on the idea of combining digital words with physical realities in different ways. It encompasses a wide spectrum of hardware. It is our conviction that this technology will play a major role in the domain of low-vision. Not only this technology can be useful to design novel vision aid-systems and rehabilitation programs, but also it has the potential to revolutionize how we study the behaviour of low-vision people (controlled condition, free head, eye tracking, possibilities for large scale studies). We have launched several projects using different platforms (see sections 5.3.1 and 6.1.1). These projects require a constant interaction with psychophysicists and ophtalmologists so as to design our solutions based on patients needs and capabilities.

3.2.3. Biophysical modeling

Modeling in neuroscience has to cope with several competing objective. On one hand, describing the biological realm as close as possible, and, on the other hand, providing tractable equations at least at the descriptive level (simulation, qualitative description) and, when possible, at the mathematical level (i.e., affording a rigorous description). These objectives are rarely achieved simultaneously and most of the time one has to make compromises. In Biovision team we adopt the point of view of physicists: try to capture the phenomenological description of a biophysical mechanism, removing irrelevant details in the description, and try to have a qualitative description of equations behaviour at least at the numerical simulation level, and, when possible, get out analytic results. We insist on the quality of the model in predicting and proposing new experiments. This requires a constant interaction with neuroscientists so as to keep the model on the tracks, warning of too crude approximation, still trying to construct equations from canonical principles [1], [2], [12].

3.2.4. Methods from theoretical physics

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

For the study of large scale populations (e.g., when studying population coding) we use methods coming from statistical physics. This branch of physics gave birth to mean-field methods as well statistical methods for large population analysis. We use both of them. Mean-field methods are applied for large scale activity in the retina and in the cortex [4], [8], [30].

For the study of retina population coding we use the so-called Gibbs distribution, initially introduced by Boltzmann and Gibbs. This concept includes, but *is not limited to*, maximum entropy models [42] used by numerous authors in the context of the retina (see, e.g., [45], [47], [41], [40], [48]). These papers were restricted to a statistical description without memory neither causality: the time correlations between successive times is not considered. However, maximum entropy extends to spatio-temporal correlations as we have shown in, e.g., [2] [49], [33]. In this context, we study how the retina respond to transient stimuli (moving objects), i.e. how spatio-temporal correlations are modified when a moving object crosses the receptive fields of ganglion cells, taking into account the lateral connectivity due to amacrine cells [19], [26], [27].

CAMIN Team

3. Research Program

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

One of CAMIN's areas of expertise is **motion measurement, observation and modeling** in the context of **sensorimotor deficiencies**. The team has the capacity to design advanced protocols to explore motor control mechanisms in more or less invasive conditions in both animal and human.
Human movement can be assessed by several noninvasive means, from motion observation (MOCAP, IMU) to electrophysiological measurements (afferent ENG, EMG, see below). Our general approach is to develop solutions that are realistic in terms of clinical or home use by clinical staff and/or patients for diagnosis and assessment purposes. In doing so, we try to gain a better understanding of motor control mechanisms, including deficient ones, which in turn will give us greater insight into the basics of human motor control. Our ultimate goal is to optimally match a neuroprosthesis to the targeted sensorimotor deficiency.
The team is involved in research projects including:

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

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

3.2. Movement assistance and/or restoration

The challenges in movement restoration are: (i) improving nerve/muscle stimulation modalities and efficiency and (ii) global management of the function that is being restored in interaction with the rest of the body under voluntary control. For this, both local (muscle) and global (function) controls have to be considered.
Online modulation of ES parameters in the context of lower limb functional assistance requires the availability of information about the ongoing movement. Different levels of complexity can be considered, going from simple open-loop to complex control laws (Figure 2).

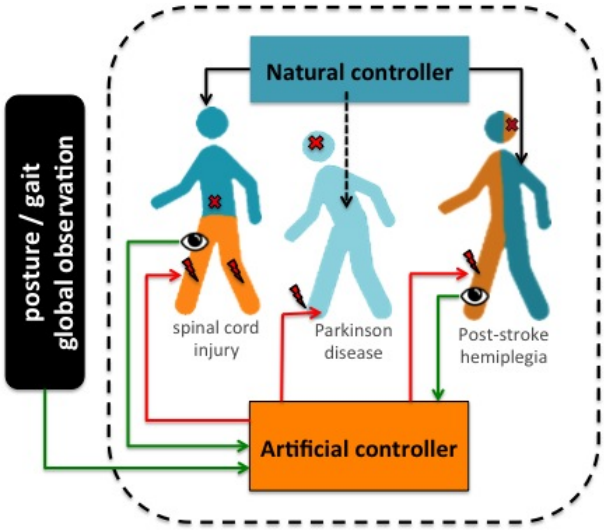


Figure 2. FES assistance should take into account the coexistence of artificial and natural controllers. Artificial controllers should integrate both global (posture/gait) and local (limb/joint) observations.

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

EPIONE Project-Team

3. Research Program

3.1. Introduction

Our research objectives are organized along 5 scientific axes:

1. Biomedical Image Analysis & Machine Learning
2. Imaging & Phenomics, Biostatistics
3. Computational Anatomy, Geometric Statistics
4. Computational Physiology & Image-Guided Therapy
5. Computational Cardiology & Image-Based Cardiac Interventions

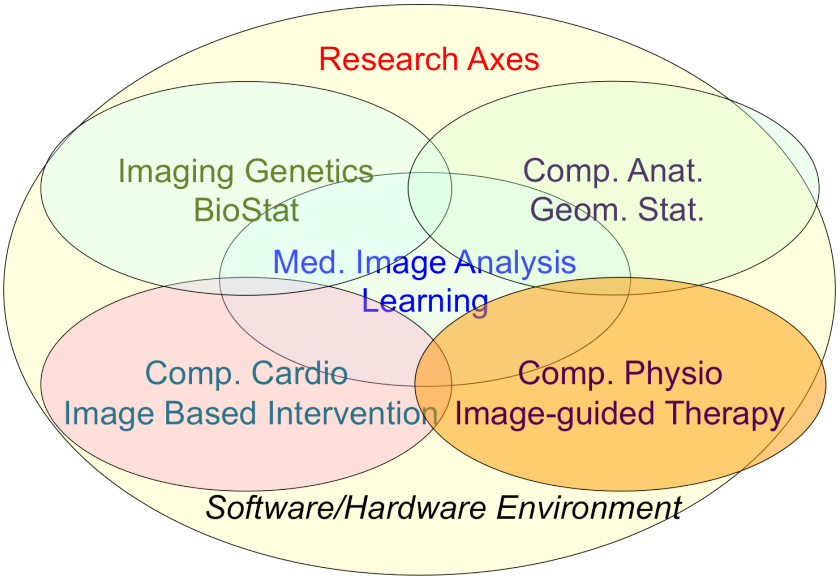


Figure 3. Epione's five main research axes

For each scientific axis, we introduce the context and the long term vision of our research.

3.2. Biomedical Image Analysis & Machine Learning

The long-term objective of biomedical image analysis is to extract, from biomedical images, pertinent information for the construction of the e-patient and for the development of e-medicine. This relates to the development of advanced segmentation and registration of images, the extraction of image biomarkers of pathologies, the detection and classification of image abnormalities, the construction of temporal models of motion or evolution from time-series of images, etc.

A good illustration of the current state of the art and of the remaining challenges can be found in these recent publications which address for instance the extraction of quantitative biomarkers on static or time varying images, as well as image registration and deformation analysis problems. This also applies to the analysis of microscopic and multi-scale images.

In addition, the growing availability of very large databases of biomedical images, the growing power of computers and the progress of machine learning (ML) approaches have opened up new opportunities for biomedical image analysis.

This is the reason why we decided to revisit a number of biomedical image analysis problems with ML approaches, including segmentation and registration problems, automatic detection of abnormalities, prediction of a missing imaging modality, etc. Not only those ML approaches often outperform the previous state-of-the-art solutions in terms of performances (accuracy of the results, computing times), but they also tend to offer a higher flexibility like the possibility to be transferred from one problem to another one with a similar framework. However, even when successful, ML approaches tend to suffer from a lack of explanatory power, which is particularly annoying for medical applications. We also plan to work on methods that can interpret the results of the ML algorithms that we develop.

- **Revisiting Segmentation problems with Machine Learning:** Through a partnership with Microsoft Research in Cambridge (UK), we are studying new segmentation methods based on deep learning with *weakly annotated* data. In effect, a complete segmentation ground truth is costly to collect in medical image analysis, as it requires the tedious task of contouring regions of interest and their validation by an expert. On the other hand, the label "presence" or "absence" of a lesion for instance (weak annotation) can be obtained at a much lower cost.

  We also plan to explore the application of deep learning methods to the fast segmentation of static or deformable organs. For instance we plan to use deep learning methods for the 3D consistent segmentation of the myocardium tissue of the 2 cardiac ventricles, an important preliminary step to mesh the cardiac muscle for computational anatomy, physiology and cardiology projects.
- **Revisiting Registration problems with Machine Learning:** We are studying, through a partnership with Siemens (Princeton), the possibility to apply robust non-rigid registration through agent-based action learning. We propose a decision process where the objective simplifies to iteratively finding the strategically next best step. An artificial agent is driven to solve the task of non-rigid registration through exploring the parametric space of a statistical deformation model built from training data. Since it is difficult to extract trustworthy ground-truth deformation fields we propose a training scheme with synthetically deformed cases and few real inter-subject cases.
- **Prediction of an imaging modality from other imaging modalities with machine learning:** Through a partnership with the Brain and Stem Institute in Paris, we plan to develop deep learning approaches to quantify some brain alterations currently measured by an invasive nuclear medicine imaging modality (PET imaging with specific tracers), directly from a multi-sequence acquisition of a non-invasive imaging modality (MRI). This requires innovative approaches taking into account the relatively small size of the ground truth database (patients having undergone both PET and MR Image acquisitions) and exploiting the a priori knowledge on the brain anatomy. We believe that this approach could apply to other image prediction problems in the longer term.
- **Prediction of cardiac pathologies with machine learning and image simulation:** Following the important work on cardiac image simulation done during the ERC project MedYMA, we are currently able to simulate time-series of images of various cardiac pathologies for which we can vary the parameters of a generative electro-mechanical model. We plan to develop new deep learning methods exploiting both the *shape* and *motion* phenotypes present in the time-series of images to detect and characterize a number of cardiac pathologies, including subtle assynchronies, local ischemia or infarcts.
- **Measuring Brain, Cognition, Behaviour:** We developed a collaborative project MNC3 which is supported by the excellence initiative Idex UCA^{Jedi}. This project gathers partners from Inria, Nice Hospitals (physicians), Nice University, and IPMC (biologists). The goal is to provide a joint analysis of heterogenous data collected on patients with neurological and psychiatric diseases. Those data include medical imaging (mainly MRI), activity (measured by connected wrists or video or microphones), biology/genomics, and clinical information. We want to show the increase in the statistical power of a joint analysis of the data to classify a pathology and to quantify its evolution.

In addition to these mid-term goals, we have applied to two important projects with local clinicians. A project on "Lung cancer", headed by anatomopathologist P. Hofman, to better exploit the joint information coming from imaging and circulating tumoral cells (in collaboration with Median Tech company); and a project "Cluster headache", headed by neurosurgeon D. Fontaine, to better integrate and exploit information coming from imaging, genetics and clinic (in collaboration with Inria Team Athena).

3.3. Imaging & Phenomics, Biostatistics

The human phenotype is associated with a multitude of heterogeneous biomarkers quantified by imaging, clinical and biological measurements, reflecting the biological and patho-physiological processes governing the human body, and essentially linked to the underlying individual genotype.In order to deepen our understanding of these complex relationships and better identify pathological traits in individuals and clinical groups, a long-term objective of e-medicine is therefore to develop the tools for the joint analysis of this heterogeneous information, termed *Phenomics*, within the unified modeling setting of the e-patient.

Ongoing research efforts aim at investigating optimal approaches at the crossroad between biomedical imaging and bioinformatics to exploit this diverse information. This is an exciting and promising research avenue, fostered by the recent availability of large amounts of data from joint imaging and biological studies (such as the UK biobank [0], ENIGMA [0], ADNI [0],...). However, we currently face important methodological challenges, which limit the ability in detecting and understanding meaningful associations between phenotype and biological information.

To date the most common approach to the analysis of the joint variation between the structure and function of organs represented in medical images, and the classical -omics modalities from biology, such as genomics or lipidomics, is essentially based on the massive univariate statistical testing of single candidate features out of the many available. This is for example the case of genome-wide association studies (GWAS) aimed at identifying statistically significant effects in pools consisting of up to millions of genetics variants. Such approaches have known limitations such as multiple comparison problems, leading to underpowered discoveries of significant associations, and usually explain a rather limited amount of data variance. Although more sophisticated machine learning approaches have been proposed, the reliability and generalization of multivariate methods is currently hampered by the low sample size relatively to the usually large dimension of the parameters space.

To address these issues this research axis investigates novel methods for the integration of this heterogeneous information within a parsimonious and unified multivariate modeling framework. The cornerstone of the project consists in achieving an optimal trade-off between modeling flexibility and ability to generalize on unseen data by developing statistical learning methods informed by prior information, either inspired by "mechanistic" biological processes, or accounting for specific signal properties (such as the structured information from spatio-temporal image time series). Finally, particular attention will be paid to the effective exploitation of the methods in the growing Big Data scenario, either in the meta-analysis context, or for the application in large datasets and biobanks.

- **Modeling associations between imaging, clinical, and biological data.** The essential aspect of this research axis concerns the study of data regularization strategies encoding prior knowledge, for the identification of meaningful associations between biological information and imaging phenotype data. This knowledge can be represented by specific biological mechanisms, such as the complex non-local correlation patterns of the -omics encoded in genes pathways, or by known spatio-temporal relationship of the data (such as time series of biological measurements or images). This axis is based on the interaction with research partners in clinics and biology, such as IPMC (CNRS, France), the Lenval Children's Hospital (France), and University College London (UK). This kind of prior information can be used for defining scalable and parsimonious probabilistic regression models. For example, it can provide relational graphs of data interactions that can be modelled by means of

[0] http://www.ukbiobank.ac.uk/
[0] http://enigma.ini.usc.edu/.
[0] http://adni.loni.usc.edu/

Bayesian priors, or can motivate dimensionality reduction techniques and sparse frameworks to limit the effective size of the parameter space. Concerning the clinical application, an important avenue of research will come from the study of the *reduced* representations of the -omics data currently available in clinics, by focusing on the modeling of the disease variants reported in previous genetic findings. The combination of this kind of data with the information routinely available to clinicians, such as medical images and memory tests, has a great potential for leading to improved diagnostic instruments. The translation of this research into clinical practice is carried out thanks to the ongoing collaboration with primary clinical partners such as the University Hospital of Nice (MNC3 partner, France), the Dementia Research Centre of UCL (UK), and the Geneva University Hospital (CH).

- **Learning from collections of biomedical databases.** The current research scenario is characterised by medium/small scale (typically from 50 to 1000 patients) heterogeneous datasets distributed across centres and countries. The straightforward extension of learning algorithms successfully applied to big data problems is therefore difficult, and specific strategies need to be envisioned in order to optimally exploit the available information. To address this problem, we focus on learning approaches to jointly model clinical data localized in different centres. This is an important issue emerging from recent large-scale multi-centric imaging-genetics studies in which partners can only share model parameters (e.g. regression coefficients between specific genes and imaging features), as represented for example by the ENIGMA imaging-genetics study, led by the collaborators at University of Southern California. This problem requires the development of statistical methods for *online* model estimation, in order to access data hosted in different clinical institutions by simply transmitting the model parameters, that will be in turn updated by using the local available data. This approach is extended to the definition of stochastic optimization strategies in which model parameters are optimized on local datasets, and then summarized in a meta-analysis context. Finally, this project studies strategies for aggregating the information from heterogeneous datasets, accounting for missing modalities due to different study design and protocols. The developed methodology finds important applications within the context of Big Data, for the development of effective learning strategies for massive datasets in the context of medical imaging (such as with the UK biobank), and beyond (ongoing collaboration with the Data Science team of EURECOM (France)).

3.4. Computational Anatomy, Geometric Statistics

Computational anatomy is an emerging discipline at the interface of geometry, statistics and image analysis which aims at developing algorithms to model and analyze the biological shape of tissues and organs. The goal is not only to establish generative models of organ anatomies across diseases, populations, species or ages but also to model the organ development across time (growth or aging) and to estimate their variability and link to other functional, genetic or structural information. Computational anatomy is a key component to support computational physiology and is evidently crucial for building the e-patient and to support e-medicine. Pivotal applications include the spatial normalization of subjects in neuroscience (mapping all the anatomies into a common reference system) and atlas to patient registration to map generic knowledge to patient-specific data. Our objectives will be to develop new efficient algorithmic methods to address the emerging challenges described below and to generate precise specific anatomical model in particular for the brain and the heart, but also other organs and structures (e.g. auditory system, lungs, breasts, etc.).

The objects of computational anatomy are often shapes extracted from images or images of labels (segmentation). The observed organ images can also be modeled using registration as the random diffeomorphic deformation of an unknown template (i.e. an orbit).In these cases as in many other applications, invariance properties lead us to consider that these objects belong to non-linear spaces that have a geometric structure. Thus, the mathematical foundations of computational anatomy rely on statistics on non-linear spaces.

- **Geometric Statistics** aim at studying this abstracted problem at the theoretical level. Our goal is to advance the fundamental knowledge in this area, with potential applications to new areas outside of medical imaging. Several challenges which constitute shorter term objectives in this direction are described below.

- **Large databases and longitudinal evolution:** The emergence of larger databases of anatomical images (ADNI, UK biobank) and the increasing availability of temporal evolution drives the need for efficient and scalable statistical techniques. A key issue is to understand how to construct hierarchical models in a non-linear setting.
- **Non-parametric models of variability:** Despite important successes, anatomical data also tend to exhibit a larger variability than what can be modeled with a standard multivariate unimodal Gaussian model. This raises the need for new statistical models to describe the anatomical variability like Bayesian statistics or sample-based statistical model like multi-atlas and archetypal techniques. A second objective is thus to develop efficient algorithmic methods for encoding the statistical variability into models.
- **Intelligible reduced-order models:** Last but not least, these statistical models should live in low dimensional spaces with parameters that can be interpreted by clinicians. This requires of course dimension reduction and variable selection techniques. In this process, it is also fundamental to align the selected variable to a dictionary of clinically meaningful terms (an ontology), so that the statistical model can not only be used to predict but also to explain.

3.4.1. Geometric Statistics

- **Foundations of statistical estimation on geometric spaces:** Beyond the now classical Riemannian spaces, this axis will develop the foundations of statistical estimation on affine connection spaces (e.g. Lie groups), quotient and stratified metric spaces (e.g. orbifolds and tree spaces). In addition to the curvature, one of the key problem is the introduction of singularities at the boundary of the regular strata (non-smooth and non-convex analysis).
- **Parametric and non-parametric dimension reduction methods in non-linear spaces:** The goal is to extend what is currently done with the Fréchet mean (i.e. a 0-dimensional approximation space) to higher dimensional subspaces and finally to a complete hierarchy of embedded subspaces (flags) that iteratively model the data with more and more precision. The Barycentric Subspace Analysis (BSA) generalization of principal component analysis which was recently proposed in the team will of course be a tool of choice for that. In this process, a key issue is to estimate efficiently not only the model parameters (mean point, subspace, flag) but also their uncertainty. Here, we want to quantify the influence of curvature and singularities on non-asymptotic estimation theory since we always have a finite (and often too limited) number of samples. As the mean is generally not unique in curved spaces, this also leads to consider that the results of estimation procedures should be changed from points to singular distributions. A key challenge in developing such a geometrization of statistics will not only be to unify the theory for the different geometric structures, but also to provide efficient practical algorithms to implement them.
- **Learning the geometry from the data:** Data can be efficiently approximated with locally Euclidean spaces when they are very finely sampled with respect to the curvature (big data setting). In the high dimensional low sample size (small data) setting, we believe that invariance properties are essential to reasonably interpolate and approximate. New apparently antagonistic notions like approximate invariance could be the key to this interaction between geometry and learning.

  Beyond the traditional statistical survey of the anatomical shapes that is developed in computational anatomy above, we intend to explore other application fields exhibiting geometric but non-medical data. For instance, applications can be found in Brain-Computer Interfaces (BCI), tree-spaces in phylogenetics, Quantum Physics, etc.

3.5. Computational Physiology & Image-Guided Therapy

Computational Physiology aims at developing computational models of human organ *functions*, an important component of the e-patient, with applications in e-medicine and more specifically in computer-aided prevention, diagnosis, therapy planning and therapy guidance. The focus of our research is on *descriptive* (allowing

to reproduce available observations), *discriminative* (allowing to separate two populations), and above all *predictive models* which can be personalized from patient data including medical images, biosignals, biological information and other available metadata. A key aspect of this scientific axis is therefore the coupling of biophysical models with patient data which implies that we are mostly considering models with relatively few and identifiable parameters. To this end, *data assimilation* methods aiming at estimating biophysical model parameters in order to reproduce available patient data are preferably developed as they potentially lead to predictive models suitable for therapy planning.

Previous research projects in computational physiology have led us to develop biomechanical models representing quasi-static small or large soft tissue deformations (e.g. liver or breast deformation after surgery), mechanical growth or atrophy models (e.g. simulating brain atrophy related to neurodegenerative diseases), heat transfer models (e.g. simulating radiofrequency ablation of tumors), and tumor growth models (e.g. brain or lung tumor growth).

To improve the data assimilation of biophysical models from patient data, a long term objective of our research will be to develop *joint imaging and biophysical generative models in a probabilistic framework* which simultaneously describe the appearance and function of an organ (or its pathologies) in medical images. Indeed, current approaches for the personalization of biophysical models often proceed in two separate steps. In a first stage, geometric, kinematic or functional features are first extracted from medical images. In a second stage, they are used by personalization methods to optimize model parameters in order to match the extracted features. In this process, subtle information present in the image which could be informative for biophysical models is often lost which may lead to limited personalization results. Instead, we propose to develop more integrative approaches where the extraction of image features would be performed jointly with the model parameter fitting. Those imaging and biophysical generative models should lead to a *better understanding* of the content of images, to a *better personalization* of model parameters and also *better estimates of their uncertainty*.

This improved coupling between images and model should *help solving various practical problems* driven by clinical applications. Depending on available resources, datasets, and clinical problems, we wish to develop a new expertise for the simulation of *tissue perfusion* (e.g. to capture the uptake of contrast agent or radioactive tracers), or *blood flow in medium / small vessels* (e.g. to capture the transport of drugs or radioactive materials in interventional radiology).

- **Reduced Computational Biophysical Models**. Clinical constraint and uncertainty estimation inevitably lead to the requirement of relatively fast computation of biophysical models. In addition to hardware acceleration (GPU, multithreading) we will explore various ways to accelerate the computation of models through intrusive (e.g. proper orthogonal decomposition, computation of condensed stiffness matrices in non-linear mechanics) or non intrusive methods (e.g. polynomial chaos expansion, Gaussian processes).
- **Uncertainty estimation of Biophysical Models**. We will pursue our research on this topic by developing Bayesian methods to estimate the posterior probability of model parameters, initial and boundary conditions from image features or image voxels. Such approaches rely on the definition of relevant likelihood terms relating the model state variables to the observable quantities in images. When possible joint imaging and biophysical generative models will be developed to avoid to rely on intermediate image features. Approximate inference of uncertainty will be estimated through Variational Bayes approaches whose accuracy will be evaluated through a comparison with stochastic sampling methods (e.g. MCMC). Through this uncertainty estimation, we also aim at developing a reliable framework to select the most sensitive and discriminative parameters of a given model but also to select the biophysical model best suited to solve a given problem (e.g. prediction of therapy outcome).
- **High Order Finite Element Modeling**. Soft tissue biomechanical models have until now been formulated as linear elastic or hyperelastic materials discretized as linear tetrahedra finite elements. While being very generic, those elements are known to suffer from numerical locking for nearly

incompressible materials and lead to poor estimate of stress field. We will develop efficient implementation and assembly methods using high order tetrahedral (and possibly hexahedral) elements. To maintain the number of nodes relatively low while keeping a good accuracy, we intend to develop elements of adaptive degree (p-refinement) driven by local error indices. Solution for meshing surfaces or volumes with curved high order elements will be developed in collaboration with the Titane and Aromath Inria teams.

- **Clinical Applications**. We plan to develop new applications of therapy planning and therapy guidance through existing or emerging collaborations related to the following problems : breast reconstruction following insertion of breast implants (with Anatoscope), planning of cochlear electrodes implantation (with CHU Nice and Oticon Medical), lung deformation following COPD or pulmonary fibrosis (with CHU Nice), echography based elastometry (with CHU Nice).

3.6. Computational Cardiology & Image-Based Cardiac Interventions

Computational Cardiology has been an active research topic within the Computational Anatomy and Computational Physiology axes of the previous Asclepios project, leading to the development of personalized computational models of the heart designed to help characterizing the cardiac function and predict the effect of some device therapies like cardiac resynchronisation or tissue ablation. This axis of research has now gained a lot of maturity and a critical mass of involved scientists to justify an individualized research axis of the new project Epione, while maintaining many constructive interactions with the 4 other research axes of the project. This will develop all the cardiovascular aspects of the e-patient for cardiac e-medicine.

The new challenges we want to address in computational cardiology are related to the introduction of new levels of modeling and to new clinical and biological applications. They also integrate the presence of new sources of measurements and the potential access to very large multimodal databases of images and measurements at various spatial and temporal scales.

Our goal will be to combine two complementary computational approaches: *machine learning* and *biophysical modelling*. This research axis will leverage on the added value of such a combination. Also we will refine our biophysical modeling by the introduction of a pharmacokinetics/pharmacodynamics (PK/PD) component able to describe the effect of a drug on the cardiac function. This will come in complement to the current geometric, electrical, mechanical and hemodynamic components of our biophysical model of the heart. We will also carefully model the uncertainty in our modeling, and try to provide algorithms fast enough to allow future clinical translation.

- Physics of Ultrasound Images for Probe Design: we will design a digital phantom of the human torso in order to help the design of echocardiographic probes. This will be done in collaboration with GE Healthcare whose excellence centre for cardiac ultrasound probes is located in Sophia Antipolis.
- Cardiac Pharmacodynamics for Drug Personalisation: we will add to our biophysical cardiac model a pharmacodynamics model, coupled with a pharmacokinetics model and a personalisation framework in order to help the adjustment of drug therapy to a given patient. This will be done in collaboration with ExactCure, a start up company specialised on this topic.
- New Imaging Modality Coupling MRI and Electrodes: we will use our fast models in order to regularize the ill-posed inverse problem of cardiac electrocardiography in order to estimate cardiac electrical activity from body surface potentials. This will be done within the ERC Starting Grant ECSTATIC coordinated by Hubert Cochet from the IHU Liryc, Bordeaux.
- Cardiac Imaging during Exercise: a particular aspect of the cardiac function is its constant adaptation to satisfy the needs of the human body. This dynamic aspect provides important information on the cardiac function but is challenging to measure. We will set up exercise protocols with Nice University Hospital and STAPS in order to model and quantify such an adaptation of the cardiac function.
- Sudden Cardiac Death is the cause of important mortality (300 000 per year in Europe, same in US) and it is difficult to identify people at risk. Based on a large multi-centric database of images, we will learn the image features correlated with a high risk of arrhythmia, with the IHU Liryc.

- Personalising models from connected objects: with the Internet of Things and the pletora of sensors available today, the cardiac function can be monitored almost continuously. Such new data open up possibilities for novel methods and tools for diagnosis, prognosis and therapy.

GALEN-POST Team

3. Research Program

3.1. Shape, Grouping and Recognition

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

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

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

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

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

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

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

3.2. Machine Learning & Structured Prediction

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

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

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

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

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

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

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

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

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

3.3. Self-Paced Learning with Missing Information

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

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

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

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

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

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

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

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

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

3.4. Discrete Biomedical Image Perception

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

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

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

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

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

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

MATHNEURO Team

3. Research Program

3.1. Neural networks dynamics

The study of neural networks is certainly motivated by the long term goal to understand how brain is working. But, beyond the comprehension of brain or even of simpler neural systems in less evolved animals, there is also the desire to exhibit general mechanisms or principles at work in the nervous system. One possible strategy is to propose mathematical models of neural activity, at different space and time scales, depending on the type of phenomena under consideration. However, beyond the mere proposal of new models, which can rapidly result in a plethora, there is also a need to understand some fundamental keys ruling the behaviour of neural networks, and, from this, to extract new ideas that can be tested in real experiments. Therefore, there is a need to make a thorough analysis of these models. An efficient approach, developed in our team, consists of analysing neural networks as dynamical systems. This allows to address several issues. A first, natural issue is to ask about the (generic) dynamics exhibited by the system when control parameters vary. This naturally leads to analyse the bifurcations [51] [52] occurring in the network and which phenomenological parameters control these bifurcations. Another issue concerns the interplay between neuron dynamics and synaptic network structure.

3.2. Mean-field and stochastic approaches

Modeling neural activity at scales integrating the effect of thousands of neurons is of central importance for several reasons. First, most imaging techniques are not able to measure individual neuron activity (microscopic scale), but are instead measuring mesoscopic effects resulting from the activity of several hundreds to several hundreds of thousands of neurons. Second, anatomical data recorded in the cortex reveal the existence of structures, such as the cortical columns, with a diameter of about $50\mu m$ to $1mm$, containing of the order of one hundred to one hundred thousand neurons belonging to a few different species. The description of this collective dynamics requires models which are different from individual neurons models. In particular, when the number of neurons is large enough averaging effects appear, and the collective dynamics is well described by an effective mean-field, summarizing the effect of the interactions of a neuron with the other neurons, and depending on a few effective control parameters. This vision, inherited from statistical physics requires that the space scale be large enough to include a large number of microscopic components (here neurons) and small enough so that the region considered is homogeneous.
Our group is developing mathematical and numerical methods allowing on one hand to produce dynamic mean-field equations from the physiological characteristics of neural structure (neurons type, synapse type and anatomical connectivity between neurons populations), and on the other so simulate these equations; see Figure 1 . These methods use tools from advanced probability theory such as the theory of Large Deviations [39] and the study of interacting diffusions [3].

3.3. Neural fields

Neural fields are a phenomenological way of describing the activity of population of neurons by delayed integro-differential equations. This continuous approximation turns out to be very useful to model large brain areas such as those involved in visual perception. The mathematical properties of these equations and their solutions are still imperfectly known, in particular in the presence of delays, different time scales and noise.
Our group is developing mathematical and numerical methods for analysing these equations. These methods are based upon techniques from mathematical functional analysis, bifurcation theory [9], [53], equivariant bifurcation analysis, delay equations, and stochastic partial differential equations. We have been able to characterize the solutions of these neural fields equations and their bifurcations, apply and expand the theory to account for such perceptual phenomena as edge, texture [31], and motion perception. We have also developed a theory of the delayed neural fields equations, in particular in the case of constant delays and propagation delays that must be taken into account when attempting to model large size cortical areas [11], [54]. This theory is based on center manifold and normal forms ideas [10].

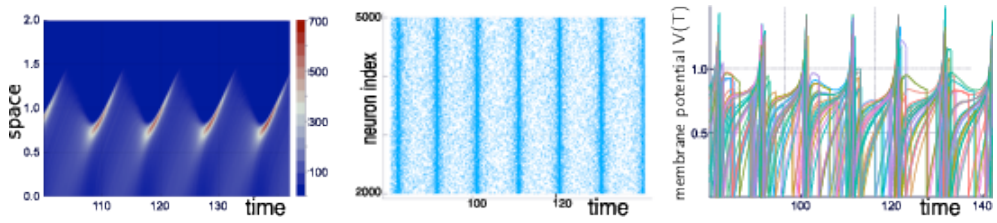


Figure 1. Simulations of the quasi-synchronous state of a stochastic neural network with $N = 5000$ neurons. Left: empirical distribution of membrane potential as a function (t, v). Middle: (raster plot) spiking times as a function of neuron index and time. Right: several membrane potentials $v_i(t)$ as a function of time for $i \in [1, 100]$. Simulated with the Julia Package PDMP.jl from [12]. This figure has been slightly modified from [7].

3.4. Slow-fast dynamics in neuronal models

Neuronal rhythms typically display many different timescales, therefore it is important to incorporate this slow-fast aspect in models. We are interested in this modeling paradigm where slow-fast point models, using Ordinary Differential Equations (ODEs), are investigated in terms of their bifurcation structure and the patterns of oscillatory solutions that they can produce. To insight into the dynamics of such systems, we use a mix of theoretical techniques — such as geometric desingularisation and centre manifold reduction [44] — and numerical methods such as pseudo-arclength continuation [36]. We are interested in families of complex oscillations generated by both mathematical and biophysical models of neurons. In particular, so-called *mixed-mode oscillations (MMOs)* [5], [34], [43], which represent an alternation between subthreshold and spiking behaviour, and *bursting oscillations* [35], [41], also corresponding to experimentally observed behaviour [32]; see Figure 2 . We are working on extending these results to spatially-extended neural models [2].

3.5. Modeling neuronal excitability

Excitability refers to the all-or-none property of neurons [38], [42]. That is, the ability to respond nonlinearly to an input with a dramatic change of response from "none" — no response except a small perturbation that returns to equilibrium — to "all" — large response with the generation of an action potential or spike before the neuron returns to equilibrium. The return to equilibrium may also be an oscillatory motion of small amplitude; in this case, one speaks of resonator neurons as opposed to integrator neurons. The combination of a spike followed by subthreshold oscillations is then often referred to as mixed-mode oscillations (MMOs) [34]. Slow-fast ODE models of dimension at least three are well capable of reproducing such complex neural oscillations. Part of our research expertise is to analyse the possible transitions between different complex oscillatory patterns of this sort upon input change and, in mathematical terms, this corresponds to understanding the bifurcation structure of the model. Furthermore, the shape of time series of this sort with a given oscillatory pattern can be analysed within the mathematical framework of dynamic bifurcations; see the section on slow-fast dynamics in Neuronal Models. The main example of abnormal neuronal excitability is hyperexcitability and it is important to understand the biological factors which lead to such excess of excitability and to identify (both in detailed biophysical models and reduced phenomenological ones) the mathematical structures leading to these anomalies. Hyperexcitability is one important trigger for pathological brain states related to various diseases such as chronic migraine, epilepsy or even Alzheimer's Disease. A central central axis of research within our group is to revisit models of such pathological scenarios, in relation with a combination of advanced mathematical tools and in partnership with biological labs.

3.6. Synaptic Plasticity

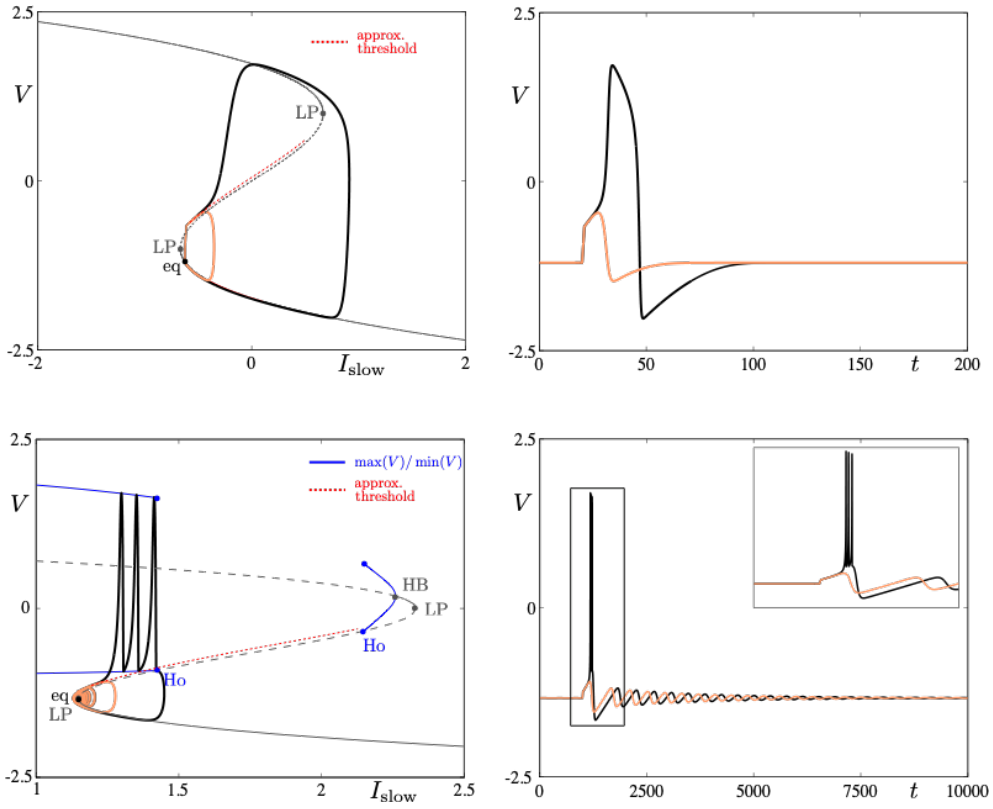


Figure 2. Excitability threshold as slow manifolds in a simple spiking model, namely the FitzHugh-Nagumo model, (top panels) and in a simple bursting model, namely the Hindmarsh-Rose model (bottom panels). This figure is unpublished.

Neural networks show amazing abilities to evolve and adapt, and to store and process information. These capabilities are mainly conditioned by plasticity mechanisms, and especially synaptic plasticity, inducing a mutual coupling between network structure and neuron dynamics. Synaptic plasticity occurs at many levels of organization and time scales in the nervous system [30]. It is of course involved in memory and learning mechanisms, but it also alters excitability of brain areas and regulates behavioral states (e.g., transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge.
Our group is developing mathematical and numerical methods to analyse this mutual interaction. On the one hand, we have shown that plasticity mechanisms [4], [8], Hebbian-like or STDP, have strong effects on neuron dynamics complexity, such as synaptic and propagation delays [11], dynamics complexity reduction, and spike statistics.

MIMESIS Team

3. Research Program

3.1. Real Time Patient-Specific Computational Models

The principal objective of this challenge is to improve, at the numerical level, the efficiency, robustness, and quality of the simulations (see Fig. 2). To reach these goals, we will investigate novel finite element techniques able to cope with complex, potentially ill-defined input data. After developing Smoothed FEM for real-time simulations, we are developing meshless techniques and immersed boundary methods. The first one is also well suited for topological changes, which we sometimes need to account for in our simulations. The second is expected to lead to more stable, and numerically efficient, formulations of the finite element method.

We will also propose numerical techniques such as domain decomposition and model order reduction, to handle real-time computation on more complex geometries or constitutive models. Boundary conditions are know to also play an important role in the solution of such problems. Therefore we are developing solutions to both identify and model the interactions that take place between the structure of interest and its anatomical environment.

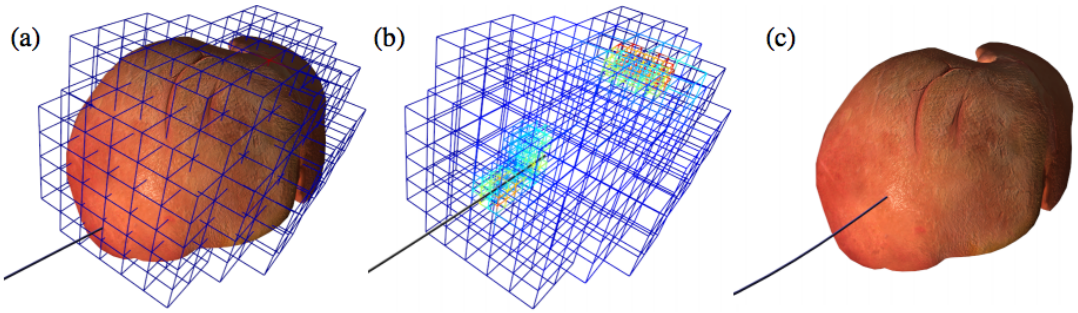


Figure 2. (a) Simulation of needle insertion in a liver; (b) Using dynamic mesh refinement scheme driven by error estimate; (c) Visual depiction. The simulation runs at 22 Hz using a PC with 4 GHz CPU.

3.2. Data-driven Simulation

Data-driven simulation has been a recent area of research in our team (see Fig. 3). We have demonstrated that it has the potential to bridge the gap between medical imaging and clinical routine by adapting pre- operative data to the time of the procedure. In the areas of non-rigid registration and augmented reality during surgery, we have demonstrated the benefit of our physics-based approaches with several key publications in major conferences (MICCAI, CVPR, IPCAI) and awards (best paper [25] at ISMAR 2013, second best paper [26] at IPCAI 2014).

We have continued this work with an **emphasis on robustness to uncertainty and outliers** in the information extracted in real-time from image data, as well as real- time parameter estimation. This is currently done by **combining Bayesian methods with advanced physics-based methods** to handle uncertainties in image-driven simulations (MICCAI 2017, CVCS 2018).

Finally, Bayesian or similar methods require to perform a large amount of simulations to sample the domain space, even when using efficient methods such as Reduced Order Unscented Kalman Filters. For this reason, we are investigating the use of neural networks to perform predictions instead of using full numerical simulations. Our latest paper [5] at MICCAI 2018 shows it is possible to **teach a neural network from numerical simulations** and **predict, with high accuracy, the relationship between an image** of the anatomy and **the associated force**.

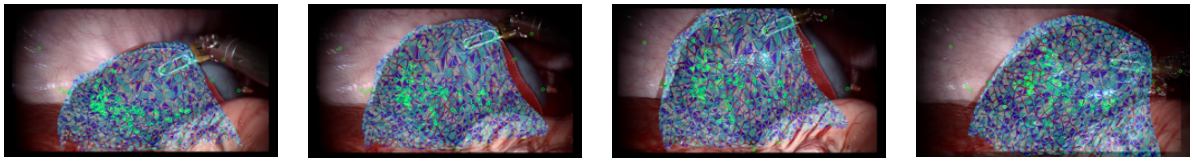


Figure 3. Real-time deformation of a virtual liver according to tissue motion tracked in laparoscopic images.

MNEMOSYNE Project-Team

3. Research Program

3.1. Integrative and Cognitive Neuroscience

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

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

This domain is today very active and stimulating because it proposes, of course at the price of simplifications, global views of cerebral functioning and more local hypotheses on the role of subsets of neuronal structures in cognition. In the global approaches, the integration of data from experimental psychology and clinical studies leads to an overview of the brain as a set of interacting memories, each devoted to a specific kind of information processing [49]. 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 [46], 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 [51], [40]. During the recent years this domain has gained an increasing interest in neuroscience and has become an essential tool for scientific developments in most fields in neuroscience, from the molecule to the system. In this view, all the objectives of our team can be described as possible progresses in computational neuroscience. Accordingly, it can be underlined that the systemic view that we promote can offer original contributions in the sense that, whereas most classical models in computational neuroscience focus on the better understanding of the structure/function relationship for isolated specific structures, we aim at exploring synergies between structures. Consequently, we target interfaces and interplay between heterogenous modes of computing, which is rarely addressed in classical computational neuroscience.

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

Here, computational neuroscience means mimicking original computations made by the neuronal substratum and mastering their corresponding properties: computations are distributed and adaptive; they are performed without an homonculus or any central clock. Numerical schemes developed for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [37], [45] and were the basis for several contributions in our group [50], [47], [52]. 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 [38]; the associated formalism [43] can be possibly considered for local studies or for relating our results with this important domain in connectionism. Nevertheless, we remain mainly at the mesoscopic level of modeling, the level of the neuronal population, and consequently interested in the formalism developed for dynamic neural fields [35], that demonstrated a richness of behavior [39] 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 [47], [48] and their use for observing the emergence of typical cortical properties [42]. In the envisioned development of more complex architectures and interplay between structures, the exploration of mathematical properties such as stability and boundedness and the observation of emerging phenomena is one important objective. This objective is also associated with that of capitalizing our experience and promoting good practices in our software production. In summary, we think that this systemic approach also brings to computational neuroscience new case studies where heterogenous and adaptive models with various time scales and parameters have to be considered jointly to obtain a mastered substratum of computation. This is particularly critical for large scale deployments.

3.3. Machine Learning

The adaptive properties of the nervous system are certainly among its most fascinating characteristics, with a high impact on our cognitive functions. Accordingly, machine learning is a domain [44] that aims at giving such characteristics to artificial systems, using a mathematical framework (probabilities, statistics, data analysis, etc.). Some of its most famous algorithms are directly inspired from neuroscience, at different levels. Connectionist learning algorithms implement, in various neuronal architectures, weight update rules, generally derived from the hebbian rule, performing non supervised (e.g. Kohonen self-organizing maps), supervised (e.g. layered perceptrons) or associative (e.g. Hopfield recurrent network) learning. Other algorithms, not necessarily connectionist, perform other kinds of learning, like reinforcement learning. Machine learning is a very mature domain today and all these algorithms have been extensively studied, at both the theoretical and practical levels, with much success. They have also been related to many functions (in the living and artificial domains) like discrimination, categorisation, sensorimotor coordination, planning, etc. and several neuronal structures have been proposed as the substratum for these kinds of learning [41], [34]. Nevertheless, we believe that, as for previous models, machine learning algorithms remain isolated tools, whereas our systemic approach can bring original views on these problems.

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

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

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

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

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

3.4. Autonomous Robotics

Autonomous robots are not only convenient platforms to implement our algorithms; the choice of such platforms is also motivated by theories in cognitive science and neuroscience indicating that cognition emerges from interactions of the body in direct loops with the world (*embodiment of cognition* [36]). In addition to real robotic platforms, software implementations of autonomous robotic systems including components dedicated to their body and their environment will be also possibly exploited, considering that they are also a tool for studying conditions for a real autonomous learning.

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

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

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

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

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

NEUROSYS Project-Team

3. Research Program

3.1. Main Objectives

The main challenge in computational neuroscience is the high complexity of neural systems. The brain is a complex system and exhibits a hierarchy of interacting subunits. On a specific hierarchical level, such subunits evolve on a certain temporal and spatial scale. The interactions of small units on a low hierarchical level build up larger units on a higher hierarchical level evolving on a slower time scale and larger spatial scale. By virtue of the different dynamics on each hierarchical level, until today the corresponding mathematical models and data analysis techniques on each level are still distinct. Only few analysis and modeling frameworks are known which link successfully at least two hierarchical levels.
After extracting models for different description levels, they are typically applied to obtain simulated activity which is supposed to reconstruct features in experimental data. Although this approach appears straightforward, it presents various difficulties. Usually the models involve a large set of unknown parameters which determine the dynamical properties of the models. To optimally reconstruct experimental features, it is necessary to formulate an inverse problem to extract optimally such model parameters from the experimental data. Typically this is a rather difficult problem due to the low signal-to-noise ratio in experimental brain signals. Moreover, the identification of signal features to be reconstructed by the model is not obvious in most applications. Consequently an extended analysis of the experimental data is necessary to identify the interesting data features. It is important to combine such a data analysis step with the parameter extraction procedure to achieve optimal results. Such a procedure depends on the properties of the experimental data and hence has to be developed for each application separately. Machine learning approaches that attempt to mimic the brain and its cognitive processes had a lot of success in classification problems during the last decade. These hierarchical and iterative approaches use non-linear functions, which imitate neural cell responses, to communicate messages between neighboring layers. In our team, we work towards developing polysomnography-specific classifiers that might help in linking the features of particular interest for building systems for sleep signal classification with sleep mechanisms, with the accent on memory consolidation during the Rapid Eye Movement (REM) sleep phase.

3.2. Challenges

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

3.3. Research Directions

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

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

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

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

PARIETAL Project-Team

3. Research Program

3.1. Inverse problems in Neuroimaging

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

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

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

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

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

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

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

with λ_1, λ_2, η_1, $\eta_2 \geq 0$ (this formulation particularly highlights the fact that convex regularizers are norms or quasi-norms). In general, only one or two of these constraints is considered (hence is enforced with a non-zero coefficient):

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

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

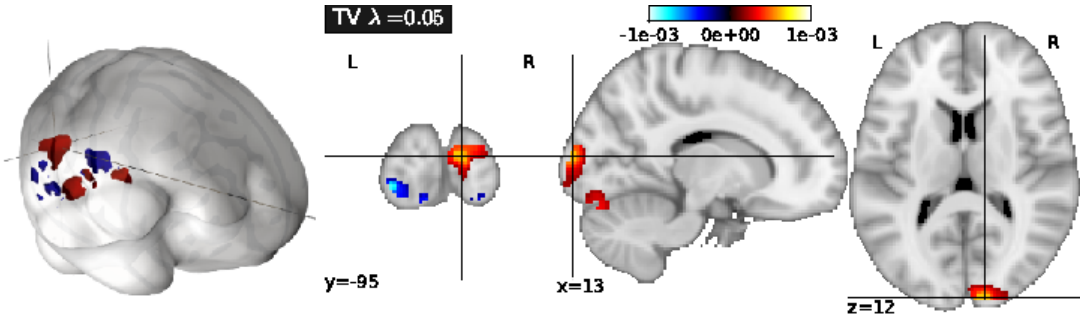


Figure 1. Example of the regularization of a brain map with total variation in an inverse problem. The problem here is to predict the spatial scale of an object presented as a stimulus, given functional neuroimaging data acquired during the presentation of an image. Learning and test are performed across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

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

This framework is easily extended by considering

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

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

then

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

3.2. Multivariate decompositions

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

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

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

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

where $(\mathbf{A}_i),\ i \in \{1..n_{features}\}$ represents the columns of $\mathbf{A}$. Ψ can be chosen such as in Eq. (2) in order to enforce smoothness and/or sparsity constraints.
The problem is not jointly convex in all the variables but each penalization given in Eq (2) yields a convex problem on $\mathbf{D}$ for $\mathbf{A}$ fixed, and conversely. This readily suggests an alternate optimization scheme, where $\mathbf{D}$ and $\mathbf{A}$ are estimated in turn, until convergence to a local optimum of the criterion. As in PCA, the extracted components can be ranked according to the amount of fitted variance. Importantly, also, estimated PCA models can be interpreted as a probabilistic model of the data, assuming a high-dimensional Gaussian distribution (probabilistic PCA).

Ultimately, the main limitations to these algorithms is the cost due to the memory requirements: holding datasets with large dimension and large number of samples (as in recent neuroimaging cohorts) leads to inefficient computation. To solve this issue, online methods are particularly attractive [24].

3.3. Covariance estimation

Another important estimation problem stems from the general issue of learning the relationship between sets of variables, in particular their covariance. Covariance learning is essential to model the dependence of these variables when they are used in a multivariate model, for instance to study potential interactions among them and with other variables. Covariance learning is necessary to model latent interactions in high-dimensional observation spaces, e.g. when considering multiple contrasts or functional connectivity data.
The difficulties are two-fold: on the one hand, there is a shortage of data to learn a good covariance model from an individual subject, and on the other hand, subject-to-subject variability poses a serious challenge to the use of multi-subject data. While the covariance structure may vary from population to population, or depending on the input data (activation versus spontaneous activity), assuming some shared structure across problems, such as their sparsity pattern, is important in order to obtain correct estimates from noisy data. Some of the most important models are:

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

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

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

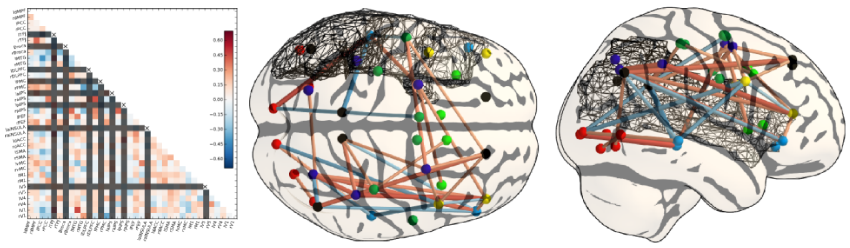


Figure 2. Example of functional connectivity analysis: The correlation matrix describing brain functional connectivity in a post-stroke patient (lesion volume outlined as a mesh) is compared to a group of control subjects. Some edges of the graphical model show a significant difference, but the statistical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

VISAGES Project-Team

3. Research Program

3.1. Research Program

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

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

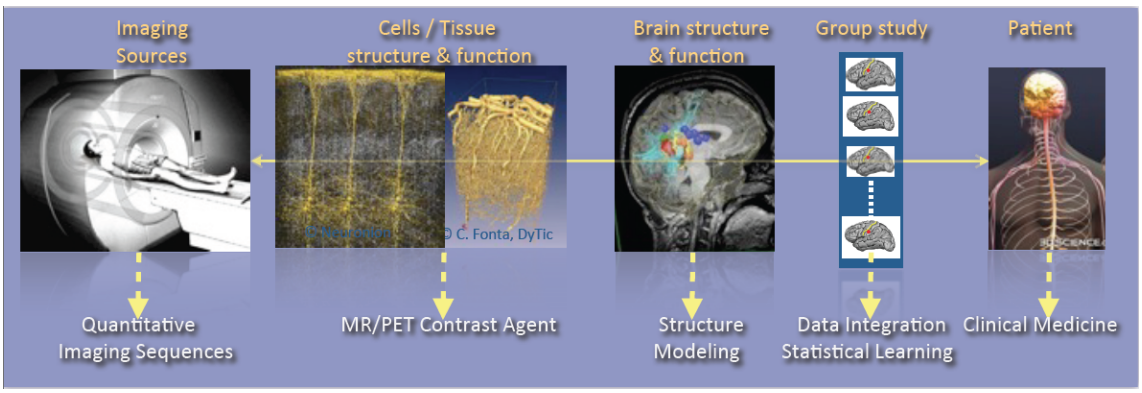


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales: from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions,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

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

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

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

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

AIRSEA Project-Team

3. Research Program

3.1. Introduction

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

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

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

3.2. Modeling for oceanic and atmospheric flows

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

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

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

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

The aim of our work is to revisit and enrich several aspects of the representation of air-sea interactions in CMs, paying special attention to their overall consistency with appropriate mathematical tools. We intend to work consistently on the physics and numerics. Using the theoretical framework of global-in-time Schwarz methods, our aim is to analyze the mathematical formulation of the parameterizations in a coupling perspective. From this study, we expect improved predictability in coupled models (this aspect will be studied using techniques described in 3.4). Complementary work on space-time nonconformities and acceleration of convergence of Schwarz-like iterative methods (see 6.1.1) are also conducted.

3.3. Model reduction / multiscale algorithms

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

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

Model reduction Reduction methods can be deterministic (proper orthogonal decomposition, other reduced bases) or stochastic (polynomial chaos, Gaussian processes, kriging), and both fields of research are very active. Choosing one method over another strongly depends on the targeted application, which can be as varied as real-time computation, sensitivity analysis (see e.g., section 6.3.1) or optimisation for parameter estimation (see below).

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

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

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

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

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

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

3.4. Dealing with uncertainties

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

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

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

Parameter estimation Advanced parameter estimation methods are barely used in ocean, atmosphere and coupled systems, mostly due to a difficulty of deriving adequate response functions, a lack of knowledge of these methods in the ocean-atmosphere community, and also to the huge associated computing costs. In the presence of strong uncertainties on the model but also on parameter values, simulation and inference are closely associated. Filtering for data assimilation and Approximate Bayesian Computation (ABC) are two examples of such association.
Stochastic approach can be compared with the deterministic approach, which allows to determine the sensitivity of the flow to parameters and optimize their values relying on data assimilation. This approach is already shown to be capable of selecting a reduced space of the most influent parameters in the local parameter space and to adapt their values in view of correcting errors committed by the numerical approximation. This approach assumes the use of automatic differentiation of the source code with respect to the model parameters, and optimization of the obtained raw code.

AIRSEA assembles all the required expertise to tackle these difficulties. As mentioned previously, the choice of parameterization schemes and their tuning has a significant impact on the result of model simulations. Our research will focus on parameter estimation for parameterized Partial Differential Equations (PDEs) and also for parameterized Stochastic Differential Equations (SDEs). Deterministic approaches are based on optimal control methods and are local in the parameter space (i.e., the result depends on the starting point of the estimation) but thanks to adjoint methods they can cope with a large number of unknowns that can also vary in space and time. Multiscale optimization techniques as described in 6.2 will be one of the tools used. This in turn can be used either to propose a better (and smaller) parameter set or as a criterion for discriminating parameterization schemes. Statistical methods are global in the parameter state but may suffer from the curse of dimensionality. However, the notion of parameter can also be extended to functional parameters. We may consider as parameter a functional entity such as a boundary condition on time, or a probability density function in a stationary regime. For these purposes, non-parametric estimation will also be considered as an alternative.

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

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

Extreme events are rare, which means there is little data available to make inferences of risk measures. Risk assessment based on observation therefore relies on multivariate extreme value theory. Interacting particle systems for the analysis of rare events is commonly used in the community of computer experiments. An open question is the pertinence of such tools for the evaluation of environmental risk.
Most numerical models are unable to accurately reproduce extreme events. There is therefore a real need to develop efficient assimilation methods for the coupling of numerical models and extreme data.

3.5. High performance computing

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

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

Parallel in time methods are iterative methods that may require a large number of iteration before convergence. Our first focus will be on the convergence analysis of parallel in time (Parareal / Schwarz) methods for the equation systems of oceanic and atmospheric models. Our second objective will be on the construction of fast (approximate) integrators for these systems. This part is naturally linked to the model reduction methods of section (6.2.1). Fast approximate integrators are required both in the Schwarz algorithm (where a first guess of the boundary conditions is required) and in the Parareal algorithm (where the fast integrator is used to connect the different time windows). Our main application of these methods will be on climate (i.e., very long time) simulations. Our second application of parallel in time methods will be in the context of optimization methods. In fact, one of the major drawbacks of the optimal control techniques used in 3.4 is a lack of intrinsic parallelism in comparison with ensemble methods. Here, parallel in time methods also offer ways to better efficiency. The mathematical key point is centered on how to efficiently couple two iterative methods (i.e., parallel in time and optimization methods).

ANGE Project-Team

3. Research Program

3.1. Overview

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

The difficulties arising in gravity driven flow studies are threefold.

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

3.2. Modelling and analysis

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

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

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

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

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

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

3.2.1. Multilayer approach

As for the first shallow water assumption, *multi-layer* systems were proposed to describe the flow as a superposition of Saint-Venant type systems [30], [33], [34]. Even if this approach has provided interesting results, layers are considered separate and non-miscible fluids, which implies strong limitations. That is why we proposed a slightly different approach [31], [32] based on a Galerkin type decomposition along the vertical axis of all variables and leading, both for the model and its discretisation, to more accurate results.

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

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

3.2.2. Non-hydrostatic models

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

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

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

3.2.3. Multi-physics modelling

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

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

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

3.2.4. Data assimilation and inverse modelling

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

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

3.3. Numerical analysis

3.3.1. Non-hydrostatic scheme

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

3.3.2. Space decomposition and adaptive scheme

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

3.3.3. Asymptotic-Preserving scheme for source terms

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

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

3.3.4. Multi-physics models

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

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

3.3.5. Optimisation

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

CASTOR Project-Team

3. Research Program

3.1. Plasma Physics

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

The main reseach topics are:

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

COFFEE Project-Team

3. Research Program

3.1. Research Program

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

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

To be more specific:

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

We can distinguish the following fields of expertise

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

FLUMINANCE Project-Team

3. Research Program

3.1. Estimation of fluid characteristic features from images

The measurement of fluid representative features such as vector fields, potential functions or vorticity maps, enables physicists to have better understanding of experimental or geophysical fluid flows. Such measurements date back to one century and more but became an intensive subject of research since the emergence of correlation techniques [50] to track fluid movements in pairs of images of a particles laden fluid or by the way of clouds photometric pattern identification in meteorological images. In computer vision, the estimation of the projection of the apparent motion of a 3D scene onto the image plane, referred to in the literature as optical-flow, is an intensive subject of researches since the 80's and the seminal work of B. Horn and B. Schunk [61]. Unlike to dense optical flow estimators, the former approach provides techniques that supply only sparse velocity fields. These methods have demonstrated to be robust and to provide accurate measurements for flows seeded with particles. These restrictions and their inherent discrete local nature limit too much their use and prevent any evolutions of these techniques towards the devising of methods supplying physically consistent results and small scale velocity measurements. It does not authorize also the use of scalar images exploited in numerous situations to visualize flows (image showing the diffusion of a scalar such as dye, p ollutant, light index refraction, fluorescein,...). At the opposite, variational techniques enable in a well-established mathematical framework to estimate spatially continuous velocity fields, which should allow more properly to go towards the measurement of smaller motion scales. As these methods are defined through PDE's systems they allow quite naturally constraints to be included such as kinematic properties or dynamic laws governing the observed fluid flows. Besides, within this framework it is also much easier to define characteristic features estimation procedures on the basis of physically grounded data model that describes the relation linking the observed luminance function and some state variables of the observed flow. The Fluminance group has allowed a substantial progress in this direction with the design of dedicated dense estimation techniques to estimate dense fluid motion fields. See [7] for a detailed review. More recently problems related to scale measurement and uncertainty estimation have been investigated [55]. Dynamically consistent and highly robust techniques have been also proposed for the recovery of surface oceanic streams from satellite images [52]. Very recently parameter-free approaches relying on uncertainty concept has been devised [53]. This technique outperforms the state of the art.

3.2. Data assimilation and Tracking of characteristic fluid features

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

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

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

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

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

3.3. Optimization and control of fluid flows with visual servoing

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

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

issue where a control law minimizing an objective cost function (minimization of the drag, minimization of the actuators power, etc.) must be applied to the actuators [51]. Most of the works of the literature indeed comes back to open-loop control approaches [65], [60], [64] or to forcing approaches [57] with control laws acting without any feedback information on the flow actual state. In order for these methods to be operative, the model used to derive the control law must describe as accurately as possible the flow and all the eventual perturbations of the surrounding environment, which is very unlikely in real situations. In addition, as such approaches rely on a perfect model, a high computational costs is usually required. This inescapable pitfall has motivated a strong interest on model reduction. Their key advantage being that they can be specified empirically from the data and represent quite accurately, with only few modes, complex flows' dynamics. This motivates an important research axis in the Fluminance group.

3.4. Numerical models applied to hydrogeology and geophysics

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

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

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

3.5. Numerical algorithms and high performance computing

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

LEMON Team

3. Research Program

3.1. Foreword

The team has three main scientific objectives. The first is to develop new models and advanced mathematical methods for inland flow processes. The second is to investigate the derivation and use of coupled models for marine and coastal processes (mainly hydrodynamics, but not only). The third is to develop theoretical methods to be used in the mathematical models serving the first two objectives. As mentioned above, the targeted applications cover PDE models and related extreme events using a hierarchy of models of increasing complexity. LEMON members also contribute to research projects that are not in the core of the team topics and that correspond to external collaborations: they are mentioned in the fourth section below.
In every section, people involved in the project are listed in alphabetical order, except for the first one (underlined) which corresponds to the leading scientist on the corresponding objective.

3.2. Inland flow processes

3.2.1. Shallow water models with porosity

3.2.1.1. State of the Art

Simulating urban floods and free surface flows in wetlands requires considerable computational power. Two-dimensional shallow water models are needed. Capturing the relevant hydraulic detail often requires computational cell sizes smaller than one meter. For instance, meshing a complete urban area with a sufficient accuracy would require 10^6 to 10^8 cells, and simulating one second often requires several CPU seconds. This makes the use of such model for crisis management impossible. Similar issues arise when modelling wetlands and coastal lagoons, where large areas are often connected by an overwhelming number of narrow channels, obstructed by vegetation and a strongly variable bathymetry. Describing such channels with the level of detail required in a 2D model is impracticable. A new generation of models overcoming this issue has emerged over the last 20 years: porosity-based shallow water models. They are obtained by averaging the two-dimensional shallow water equations over large areas containing both water and a solid phase [44]. The size of a computational cell can be increased by a factor 10 to 50 compared to a 2D shallow water model, with CPU times reduced by 2 to 3 orders of magnitude [67]. While the research on porosity-based shallow water models has accelerated over the past decade [61], [80], [84], [56], [55], [67], [96], [97], [91], [92], a number of research issues remain pending.

3.2.1.2. Four year research objectives

The research objectives are (i) to improve the upscaling of the flux and source term models to be embedded in porosity shallow water models, (ii) to validate these models against laboratory and in situ measurements. Improving the upscaled flux and source term models for urban applications requires that description of anisotropy in porosity models be improved to account for the preferential flows induced by building and street alignment. The description of the porosity embedded in the most widespread porosity approach, the so-called Integral Porosity model [80], [58], has been shown to provide an incomplete description of the connectivity properties of the urban medium. Firstly, the governing equations are strongly mesh-dependent because of consistency issues [58]. Secondly, the flux and source term models fail to reproduce the alignment with the main street axes in a number of situations [57]. Another path for improvement concerns the upscaling of obstacle-induced drag terms in the presence of complex geometries. Recent upscaling research results obtained by the LEMON team in collaboration with Tour du Valat suggest that the effects of microtopography on the flow cannot be upscaled using "classical" equation-of-state approaches, as done in most hydraulic models. A totally different approach must be proposed. The next four years will be devoted to the development and validation of improved flux and source term closures in the presence of strongly anisotropic urban geometries

and in the presence of strongly variable topography. Validation will involve not only the comparison of porosity model outputs with refined flow simulation results, but also the validation against experimental data sets. No experimental data set allowing for a sound validation of flux closures in porosity models can be found in the literature. Laboratory experiments will be developed specifically in view of the validation of porosity models. Such experiments will be set up and carried out in collaboration with the Université Catholique de Louvain (UCL), that has an excellent track record in experimental hydraulics and the development of flow monitoring and data acquisition equipment. These activities will take place in the framework of the PoroCity Associate International Laboratory (see next paragraph).

3.2.1.3. People

Vincent Guinot, Carole Delenne, Antoine Rousseau.

3.2.1.4. External collaborations

- Tour du Valat (O. Boutron): the partnership with TdV focuses on the development and application of depth-dependent porosity models to the simulation of coastal lagoons, where the bathymetry and geometry is too complex to be represented using refined flow models.
- University of California Irvine (B. Sanders): the collaboration with UCI started in 2014 with research on the representation of urban anisotropic features in integral porosity models [67]. It has led to the development of the Dual Integral Porosity model [59]. Ongoing research focuses on improved representations of urban anisotropy in urban floods modelling.
- Université Catholique de Louvain - UCL (S. Soares-Frazão): UCL is one of the few places with experimental facilities allowing for the systematic, detailed validation of porosity models. The collaboration with UCL started in 2005 and will continue with the PoroCity Associate International Laboratory proposal. In this proposal, a four year research program is set up for the validation, development and parametrization of shallow water models with porosity.

3.2.2. Forcing

3.2.2.1. State of the Art

Reproducing optimally realistic spatio-temporal rainfall fields is of salient importance to the forcing of hydrodynamic models. This challenging task requires combining intense, usual and dry weather events. Far from being straightforward, this combination of extreme and non-extreme scenarii requires a realistic modelling of the transitions between normal and extreme periods. [72] have proposed in a univariate framework a statistical model that can serve as a generator and that takes into account low, moderate and intense precipitation. In the same vein, [93] developed a bivariate model. However, its extension to a spatial framework remains a challenge. Existing spatial precipitation stochastic generators are generally based on Gaussian spatial processes [30], [69], that are not adapted to generate extreme rainfall events. Recent advances in spatio-temporal extremes modelling based on generalized Pareto processes [48], [87] and semi-parametric simulation techniques [36] are very promising and could form the base for relevant developments in our framework.

3.2.2.2. Four year research objectives

The purpose is to develop stochastic methods for the simulation of realistic spatio-temporal processes integrating extreme events. Two steps are identified. The first one is about the simulation of extreme events and the second one concerns the combination of extreme and non extreme events in order to build complete, realistic precipitations time series. As far as the first step is concerned, a first task is to understand and to model the space-time structure of hydrological extremes such as those observed in the French Mediterranean basin, that is known for its intense rainfall events (Cevenol episodes), which have recently received increased attention. We will propose modelling approaches based on the exceedance, which allows the simulated fields to be interpreted as events. Parametric, semi-parametric and non-parametric approaches are currently under consideration. They would allow a number of scientific locks to be removed. Examples of such locks are e.g. accounting for the temporal dimension and for various dependence structures (asymptotic dependence or asymptotic independence possibly depending on the dimension and/or the distance considered). Methodological aspects are detailed in Section 3.4.1 . The second step, which is not straightforward, consists in combining different spatio-temporal simulations in order to help to ultimately develop a stochastic precipitation generator capable of producing full precipitation fields, including dry and non-extreme wet periods.

3.2.2.3. People

Gwladys Toulemonde, Carole Delenne, Vincent Guinot.

3.2.2.4. External collaborations

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

3.2.3. Parametrization of shallow water models with porosity

3.2.3.1. State of the Art

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

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

Remote sensing data are becoming widely available with high spatial and temporal resolutions. Several recent studies have shown that flood extends can be extracted from optical or radar images [51], for example: to characterize the flood dynamics of great rivers [73], to monitor temporary ponds [85], but also to calibrate hydrodynamics models and assess roughness parameters [82], [62], [95].

Upscaled models developed in LEMON (see 3.2.1) embed new parameters that reflect the statistical properties of the medium geometry. Two types of information are needed: the directional properties of the medium and its flow connectivity properties. New methods are thus to be developed to characterize such statistical properties from geographical data.

3.2.3.2. Four year research objectives

This research line consists in deriving methods and algorithms for the determination of upscaled model parameters from geodata. In developed countries, it is intended to extract information on the porosity parameters and their principal directions from National geographical survey databases. Such databases usually incorporate separate layers for roads, buildings, parking lots, yards, etc. Most of the information is stored in vector form, which can be expected to make the treatment of urban anisotropic properties easier than with a raster format. In developing countries, data is made increasingly available over the world thanks to crowdsourcing (e.g. OpenStreetMap). However, such level of detail in vector format is still not available in many countries. Moreover, vector data for the street network does not provide all the relevant information. In suburban areas, lawns, parks and other vegetated areas may also contribute to flood propagation and storage. In this context, it is intended to extract the necessary information from aerial and/or satellite images, that are widely available and the spatial resolution of which improves constantly. A research line will consist in deriving the information on street preferential orientation using textural analysis techniques. Such techniques have been used successfully in the field of agricultural pattern identification during Carole Delenne's PhD thesis [46], [77]. However, their application to the urban environment raises a number of issues. One of them is the strongly discontinuous character of the urban medium, that makes textural analysis difficult.

Moreover, in order to achieve a correct parametrization, identifying areas with homogeneous porosity properties is necessary. Algorithms identifying the shape and extension of such areas are still to be developed.

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

3.2.3.3. People

Carole Delenne, Vincent Guinot, Antoine Rousseau

3.2.3.4. External collaborations

- The methodologies concerning geographical databases in vector form will be developed in strong collaboration with C. Dieulin at HSM in the framework of the PoroCity Associate International Laboratory cited above.
- Research on topography reconstruction in wetlands begun in collaboration with J.-S. Bailly (LISAH) in 2016 [45] and will continue in the coming years.

3.3. Marine and coastal systems

3.3.1. Multi-scale ocean modelling

The expertise of LEMON in this scientific domain is more in the introduction and analysis of new boundary conditions for ocean modelling systems, that can be tested on academical home-designed test cases. This is in the core of Antoine Rousseau's contributions over the past years. The real implementation, within operational ocean models, has to be done thanks to external collaborations which have already started with LEMON (see below).

3.3.1.1. State of the Art

In physical oceanography, all operational models - regardless of the scale they apply to - are derived from the complete equations of geophysical fluid dynamics. Depending on the considered process properties (nonlinearity, scale) and the available computational power, the original equations are adapted with some simplifying hypotheses. The reader can refer to [79], [70] for a hierarchical presentation of such models.
In the nearshore area, the hydrostatic approximation that is used is most large scales models (high sea) cannot be used without a massive loss of accuracy. In particular, shallow water models are inappropriate to describe the physical processes that occur in this zone (see Figure 1). This is why Boussinesq-type models are prefered: see [68]. They embed dispersive terms that allow for shoaling and other bathymetry effects. Since the pioneering works of Green and Naghdi (see [52]), numerous theoretical and numerical studies have been delivered by the "mathematical oceanography" community, more specifically in France (see the works of Lannes, Marche, Sainte-Marie, Bresch, etc.). The corresponding numerical models (BOSZ, WaveBox) must thus be integrated in any reasonable nearshore modelling platform.

However, these models cannot simply replace all previous models everywhere in the ocean: dispersive models are useless away from the shore and it is known that wave breaking cannot be simulated using Boussinesq-type equations. Hence the need to couple these models with others. Some work has been done in this direction with a multi-level nesting using software packages such as ROMS, but to the best of our knowledge, all the "boxes" rely on the same governing equations with different grid resolutions. A real coupling between different models is a more difficult task since different models may have different mathematical properties, as shown in the work by Eric Blayo and Antoine Rousseau on shallow water modelling (see [32]).

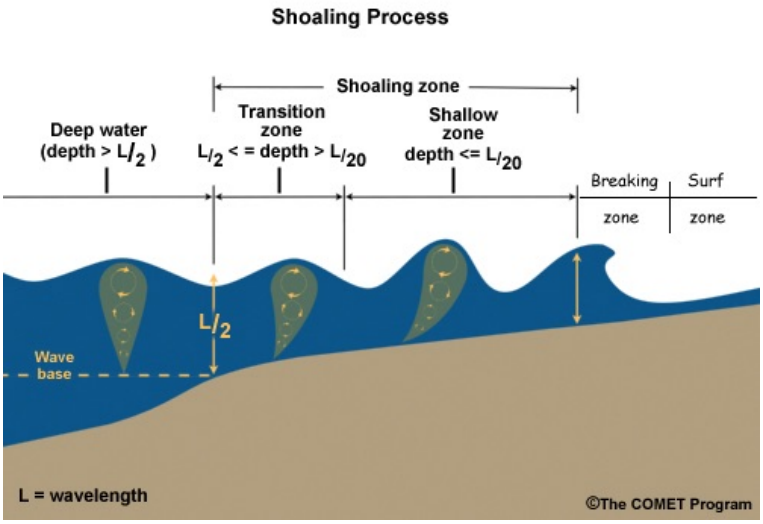


Figure 1. Deep sea, shoaling, and breaking zones.

3.3.1.2. Four year research objectives

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

- **Dispersive *vs* non-dispersive shallow water models**. As depicted in Figure 1 above, Boussinesq-type models (embedding dispersive effects) should be used in the so-called shoaling zone. The coupling with classical deep-sea / shallow water models has to be done such that all the processes in Figure 1 are correctly modelled (by different equations), with a reduced numerical cost. As a first guess, we think that Schwarz-type methods (widely used by the DDM community) could be good candidates, in particular when the interface locations are well-known. Moving interfaces (depending on the flow, the bathymetry and naturally the wind and all external forcings) is a more challenging objective that will be tackled after the first step (known interface) is achieved.
- **spectral *vs* time-domain models**. In the context of mathematical modelling and numerical simulation for the marine energy, we want to build a coupled numerical model that would be able to simulate wave propagation in domains covering both off-shore regions, where spectral models are used, and nearshore regions, better described by nonlinear dispersive (Boussinesq-type) models.
  While spectral models work with a statistical and phase-averaged description of the waves, solving the evolution of its energy spectrum, Boussinesq-type models are phase-resolving and solves nonlinear dispersive shallow water equations for physical variables (surface elevation and velocity) in the time domain. Furthermore, the time and space scales are very different: they are much larger in the case of spectral models, which justifies their use for modelling off-shore propagation over large time frames. Moreover, important small scale phenomena in nearshore areas are better captured by Boussinesq models, in which the time step is limited by the CFL condition.
  From a mathematical and modelling point of view, this task mainly consists in working on the boundary conditions of each model, managing the simultaneous use of spectral and time series data, while studying transparent boundary conditions for the models and developing domain decomposition approaches to improve the exchange of information.

3.3.1.3. People

Antoine Rousseau, Joao Guilherme Caldas Steinstraesser

3.3.1.4. External collaborations

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

3.3.2. Data-model interactions

3.3.2.1. State of the Art

An alternative to direct observations is the chaining of numerical models, which for instance represent the physic from offshore to coastal areas. Typically, output data from atmospheric and ocean circulation models are used as forcings for a wave model, which in turn feeds a littoral model. In the case of extreme events, their numerical simulation from physical models is generally unreachable. This is due to a lack of knowledge on boundary conditions and on their physical reliability for such extreme quantities. Based on numerical simulated data, an alternative is to use statistical approaches. [36] proposed such an approach. They first produced and studied a 52-year hindcast using the WW3 wave model [34], [37], [35], [88]. Then stemming from parts of the original work of [33], [53], [48], [36] proposed a semi-parametric approach which aims to simulate extreme space-time waves processes to, in turn, force a littoral hazard model. Nevertheless their approach allows only a very small number of scenarii to be simulated.

3.3.2.2. Four year research objectives

A first objective is to establish the link between the simulation approach proposed by [36] and the Pareto Processes [48]. This will allow the work of [36] to be generalized, thus opening up the possibility of generating an infinity of extreme scenarii. While continuing to favor the semi- or non-parametric approaches made possible by the access to high spatial resolution calculations, we will try to capture the strength of potentially decreasing extremal dependence when moving towards higher values, which requires the development of models that allow for so-called asymptotic independence.

3.3.2.3. People

Gwladys Toulemonde, Fátima Palacios Rodríguez, Antoine Rousseau

3.3.2.4. External collaborations

- The collaboration with Romain Chailan (IMAG, UM, CNRS) and Frédéric Bouchette (Geosciences, UM) started in 2012 during the PhD of Romain entitled Application of scientific computing and statistical analysis to address coastal hazards.
- During her post doctoral position, Fátima Palacios Rodríguez with her co-advisors will considered a generalization of the proposed simulation method by [36].

3.4. Methodological developments

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

3.4.1. Stochastic models for extreme events

3.4.1.1. State of the Art

Max-stable random fields [83], [81], [65], [41], [74] are the natural limit models for spatial maximum data and have spawned a very rich literature. An overview of typical approaches to modelling maxima is due to [43]. Physical interpretation of simulated data from such models can be discussed. An alternative to the max-stable framework are models for threshold exceedances. Processes called GPD processes, which appear as a generalization of the univariate formalism of the high thresholds exceeding a threshold based on the GPD, have been proposed [48], [87]. Strong advantages of these thresholding techniques are their capability to exploit more information from the data and explicitly model the original event data. However, the asymptotic dependence stability in these limiting processes for maximum and threshold exceedance tends to be overly restrictive when asymptotic dependence strength decreases at high levels and may ultimately vanish in the case of asymptotic independence. Such behaviours appear to be characteristic for many real-world data sets such as precipitation fields [42], [86]. This has motivated the development of more flexible dependence models such as max-mixtures of max-stable and asymptotically independent processes [94], [28] for maxima data, and Gaussian scale mixture processes [75], [64] for threshold exceedances. These models can accommodate asymptotic dependence, asymptotic independence and Gaussian dependence with a smooth transition. Extreme events also generally present a temporal dependence [89] . Developing flexible space-time models for extremes is crucial for characterizing the temporal persistence of extreme events spanning several time steps; such models are important for short-term prediction in applications such as the forecasting of wind power and for extreme event scenario generators providing inputs to impact models, for instance in hydrology and agriculture. Currently, only few models are available from the statistical literature (see for instance [39], [40], [63]) and remain difficult to interpret.

3.4.1.2. Four year research objectives

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

3.4.1.3. People

Gwladys Toulemonde, Fátima Palacios Rodríguez

3.4.1.4. External collaborations

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

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

3.4.2. Integrating heterogeneous data

3.4.2.1. State of the Art

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

The assumption is made that even hardly accessible information often exists. This stems from the increasing availability of remote-sensing data, to the crowd-sourcing of geographical databases, including the inexhaustible source of information provided by the Internet. However, information remains quite fragmented and stored in various formats: images, vector shapes, texts, etc.

This path of research begun with the Cart'Eaux project (2015-2018), that aims to produce regular and complete mapping of urban wastewater system. Contrary to drinkable water networks, the knowledge of sewer pipe location is not straightforward, even in developed countries. Over the past century, it was common practice for public service providers to install, operate and repair their networks separately [78]. Now local authorities are confronted with the task of combining data produced by different parts, having distinct formats, variable precision and granularity [38].

3.4.2.2. Four year research objectives

The overall objective of this research line is to develop methodologies to gather various types of data in the aim of producing an accurate mapping of the studied systems for hydrodynamics models.

Concerning wastewater networks, the methodology applied consists in inferring the shape of the network from a partial dataset of manhole covers that can be detected from aerial images [76]. Since manhole covers positions are expected to be known with low accuracy (positional uncertainty, detection errors), a stochastic algorithm is set up to provide a set of probable network geometries. As more information is required for hydraulic modelling than the simple mapping of the network (slopes, diameters, materials, etc.), text mining techniques such as used in [66] are particularly interesting to extract characteristics from data posted on the Web or available through governmental or specific databases. Using an appropriate keyword list, thematic entities are identified and linked to the surrounding spatial and temporal entities in order to ease the burden of data collection. It is clear at this stage that obtaining numerical values on specific pipes will be challenging. Thus, when no information is found, decision rules will be used to assign acceptable numerical values to enable the final hydraulic modelling.

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

In collaboration with J.S. Bailly (LISAH), it is intended to extend the application field of the Cart'Eaux project to rainwater collection systems, involving free surface ditches. These are particularly present in peri-urban areas and are integral part of the green corridor by playing a crucial environmental role of pollution retention and ecological continuity. Multiple-point geostatistics methods [54] will be explored, especially the Direct Sampling approach [71], efficient to simulate spatial heterogeneities by combining continuous and categorized data. When a variable is observed at a given location, the method uses it as conditional information to guide the simulation of another variable of interest in the whole spatial field.

Combining heterogeneous data for a better knowledge of studied systems raises the question of data fusion. What is the reality when contradictory information is collected from different sources? Dealing with spatial information, offset are quite frequent between different geographical data layers; pattern comparison approaches should be developed to judge whether two pieces of information represented by two elements close to each other are in reality identical, complementary, or contradictory.

3.4.2.3. People

Carole Delenne, Vincent Guinot, Antoine Rousseau, Gwladys Toulemonde

3.4.2.4. External collaborations

The Cart'Eaux project has been a lever to develop a collaboration with Berger-Levrault company and several multidisciplinary collaborations for image treatment (LIRMM), text analysis (LIRMM and TETIS) and network cartography (LISAH, IFSTTAR).

- A new project lead by N. Chahinian (HSM) has recently been funded concerning data mining and text analysis, in collaboration with linguists of URM Praxiling. Carole Delenne will have a slight implication in this project.
- A phd thesis will be submitted to the French Association of Research and Technology (ANRT) in co-funding with Berger-Levrault company concerning data fusion.
- The problematic of inferring a connected network from scarce or uncertain data is common to several research topics in LEMON such as sewage or drainage systems, urban media and wetlands. A generic methodology will be developed in collaboration with J.-S. Bailly (LISAH).

3.4.3. Numerical methods for porosity models

3.4.3.1. State of the Art

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

Firstly, the most recently developed DIP models include a momentum source term involving the divergence of the momentum fluxes [59]. This source term is not active in all situations but takes effect only when positive waves are involved [56], [57]. The consequence is a discontinuous flux tensor and discontinuous wave propagation properties. The consequences of this on the existence and uniqueness of solutions to initial value problems (especially the Riemann problem) are not known, or are the consequences on the accuracy of the numerical methods used to solve this new type of equations.

Secondly, most applications of these models involve anisotropic porosity fields [67], [80]. Such anisotropy can be modelled using 2×2 porosity tensors, with principal directions that are not aligned with those of the Riemann problems in two dimensions of space. The solution of such Riemann problems has not been investigated yet. Moreover, the governing equations not being invariant by rotation, their solution on unstructured grids is not straightforward.

Thirdly, the Riemann-based, finite volume solution of the governing equations require that the Riemann problem be solved in the presence of a porosity discontinuity. While recent work [47] has addressed the issue for the single porosity equations, similar work remains to be done for integral- and multiple porosity-based models.

3.4.3.2. Four year research objectives

The four year research objectives are the following:

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

3.4.3.3. People

Vincent Guinot

3.4.3.4. External collaborations

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

- Minh Le, Saint Venant laboratory, Chatou (France): numerical methods for shallow water flows, experience with the 2D, finite element/finte volume-based Telemac2D system.
- M.E. Vazquez-Cendon, Univ. Santiago da Compostela (Spain): finite volume methods for shallow water hydrodynamics and transport, developed Riemann solvers for the single porosity equations.

- A. Ferrari, R. Vacondio, S. Dazzi, P. Mignosa, Univ. Parma (Italy): applied mathematics, Riemann solvers for the single porosity equations.
- O. Delestre, Univ. Nice-Sophia Antipolis (France): development of numerical methods for shallow water flows (source term treatment, etc.)
- F. Benkhaldoun, Univ. Paris 13 (France): development of Riemann solvers for the porous shallow water equations.

.

3.4.4. External collaborations

3.4.4.1. Inland hydrobiological systems

3.4.4.1.1. State of the Art

Water bodies such as lakes or coastal lagoons (possibly connected to the sea) located in high human activity areas are subject to various kinds of stress such as industrial pollution, high water demand or bacterial blooms caused by freshwater over-enrichment. For obvious environmental reasons, these water resources have to be protected, hence the need to better understand and possibly control such fragile ecosystems to eventually develop decision-making tools. From a modelling point of view, they share a common feature in that they all involve interacting biological and hydrological processes. According to [49], models may be classified into two main types: “minimal dynamic models” and “complex dynamic models”. These two model types do not have the same objectives. While the former are more heuristic and rather depict the likelihood of considered processes, the latter are usually derived from fundamental laws of biochemistry or fluid dynamics. Of course, the latter necessitate much more computational resources than the former. In addition, controlling such complex systems (usually governed by PDEs) is by far more difficult that controlling the simpler ODE-driven command systems.
LEMON has already contributed both to the reduction of PDE models for the simulation of water confinement in coastal lagoons [50], [31] and to the improvement of ODE models in order to account for space-heterogeneity of bioremediation processes in water resources [29].

3.4.4.1.2. Four year research objectives

In collaboration with colleagues from the ANR-ANSWER project and colleagues from INRA, our ambition is to improve existing models of lagoon/marine ecosystems by integrating both accurate and numerically affordable coupled hydrobiological systems. A major challenge is to find an optimal trade-off between the level of detail in the description of the ecosystem and the level of complexity in terms of number of parameters (in particular regarding the governing equations for inter-species reactions). The model(s) should be able to reproduce the inter-annual variability of the observed dynamics of the ecosystem in response to meteorological forcing. This will require the adaptation of hydrodynamics equations to such time scales (reduced/upscaled models such as porosity shallow water models (see Section 3.2.1) will have to be considered) together with the coupling with the ecological models. At short time scales (i.e. the weekly time scale), accurate (but possibly CPU-consuming) 3D hydrodynamic models processes (describing thermal stratification, mixing, current velocity, sediment resuspension, wind waves...) are needed. On the longer term, it is intended to develop reduced models accounting for spatial heterogeneity.
The team will focus on two main application projects in the coming years:

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

3.4.4.1.3. People

Céline Casenave (INRA Montpellier), Antoine Rousseau, Vincent Guinot, Joseph Luis Kahn Casapia, PhD student (march 2018)

3.4.4.1.4. Collaborations

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

MAGIQUE-3D Project-Team

3. Research Program

3.1. Introduction

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

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

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

3.2. Mathematical modeling of multi-physics involving wave equations

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

The accurate numerical simulation of full wave problems in heterogeneous media is computationally intensive since it needs numerical schemes based on grids. The size of the cells depends on the propagation velocity of waves. When coupling wave problems, conversion phenomena may occur and waves with very different propagation velocity coexist. The size of the cells is then defined from the smallest velocity and in most of the real-world cases, the computational cost is crippling. Regarding existing computing capabilities, we propose to derive intermediate models which require less computational burden and provide accurate solutions for a wide-ranging class of problems including Elasto-acoustics and Electro-seismology.
When it comes to mathematical analysis, we have identified two tasks which could help us simulate realistic 3D multi-physics wave problems and which are in the scope of our savoir-faire. They are construction of approximate and multiscale models which are different tasks. The construction of approximate problems aims at deriving systems of equations which discrete formulation involves middle-sized matrices and in general, they are based on high frequency hypothesis. Multiscale models are based on a rigorous analysis involving a small parameter which does not depend on the propagation velocity necessarily.
Recently, we have conducted research on the construction of approximate models for offshore imaging. Elastic and acoustic wave equations are coupled and we investigate the idea of eliminating the computations inside water by introducing equivalent interface conditions on the sea bottom. We apply an On-Surface-Radiation-Condition (OSRC) which is obtained from the approximation of the acoustic Dirichlet-to-Neumann (DtN) operator [70], [49]. To the best of our knowledge, OSRC method has never been used for solving reduced coupling wave problems and preliminary promising results are available at [51]. We would like to investigate this technique further because we could form a battery of problems which can be solved quickly. This would provide a set of solutions which we could use as initial guess for solving inverse problems. But we are concerned with the performance of the OSRC method when wave conversions with different wavelengths occur. Anyway, the approximation of the DtN operator is not obvious when the medium is strongly heterogeneous and multiscale analysis might be more adapted. For instance, according to existing results in Acoustics and Electromagnetism for the modeling of wire antennas [61], multiscale analysis should turn out to be very efficient when the propagation medium includes well logs, fractures and faults which are very thin structures when compared to the wavelength of seismic waves. Moreover, multiscale analysis should perform well when the medium is strongly oscillating like porous media. It could thus provide an alternative to homogenization techniques which can be applied only when the medium is periodic. We thus propose to develop reduced multi-scale models by performing rigorous mathematical procedure based on regular and singular multiscale analysis. Our approach distinguishes itself from others because it focuses on the numerical representation of small structures by time-dependent problems. This could give rise to the development of new finite element methods which would combine DG approximations with XFEM (Extended Finite Element Method) which has been created for the finite element treatment of thin structures like cracks.
But Earth imaging must be more than using elasto-acoustic wave propagation. Electromagnetic waves can also be used and in collaboration with Prof. D. Pardo (Iker Basque Foundation and University of Bilbao), we conduct researches on passive imaging to probe boreholes. Passive imaging is a recent technique of imaging which uses natural electromagnetic fields as sources. These fields are generated by hydromagnetic waves propagating in the magnetosphere which transform into electromagnetic waves when they reach the ionosphere. This is a mid-frequency imaging technique which applies also to mineral and geothermal exploration, to predict seismic hazard or for groundwater monitoring. We aim at developing software package for resistivity inversion, knowing that current numerical methods are not able to manage 3D inversion. We have obtained results based on a Petrov-Galerkin approximation [46], but they are limited to 2D cases. We have thus proposed to reduce the 3D problem by using 1D semi-analytic approximation of Maxwell equations [74]. This work has just started in the framework of a PhD thesis and we hope that it will give us the possibility of imaging 3D problems.
Magique-3D would like to expand its know-how by considering electro-seismic problems which are in the scope of coupling electromagnetic waves with seismic waves. Electro-seismic waves are involved in porous media imaging which is a tricky task because it is based on the coupling of waves with very different wavelengths described by Biot equations and Maxwell equations. Biot equations govern waves in saturated porous media and they represent a complex physical phenomenon involving a slow wave which is very difficult to simulate numerically. In [68], interesting results have been obtained for the simulation of piezoelectric

sensors. They are based on a quasi-static approximation of the Maxwell model coupled with Elastodynamics. Now, we are concerned with the capability of using this model for Geophysical Imaging and we believe that the derivation and/or the analysis of suitable modelings is necessary. Collaborations with Geophysicists are thus mandatory in the prospect of using both experimental and numerical approaches. We would like to collaborate with Prof. C. Bordes and Prof. D. Brito (Laboratory of Complex Fluids and their Reservoirs, CNRS and University of Pau) who have efficient experimental devices for the propagation of electromagnetic waves inside saturated porous media [50]. This collaboration should be easy to organize since Magique-3D has a long-term experience in collaborating with geophysicists. We then believe that we will not need a lot of time to get joint results since we can use our advanced software packages Hou10ni and Montjoie and our colleagues have already obtained data. Electro-seismology is a very challenging research domain for us and we would like to enforce our collaborations with IsTerre (Institute of Earth Science, University of Grenoble) and for that topic with Prof. S. Garambois who is an expert in Electro-seismology [76], [77], [65], [66]. A joint research program could gather Geophysicists from the University of Pau and from IsTerre and Magique-3D. In particular, it would be interesting to compare simulations performed with Hou10ni, Montjoie, with the code developed by Prof. S. Garambois and to use experimental simulations for validation.

3.3. Supercomputing for Helmholtz problems

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

During the period 2010-2014, the team has worked extensively on high order discontinuous Galerkin (DG) methods. Like standard Finite Element Methods, they are elaborated with polynomial basis functions and they are very popular because they are defined locally for each element. It is thus easy to use basis polynomial functions with different degrees and this shows the perfect flexibility of the approximation in case of heterogeneous media including homogeneous parts. Indeed, low degree basis functions can be used in heterogeneous regions where a fine grid is necessary while high degree polynomials can be used for coarse elements covering homogeneous parts. In particular, Magique-3D has developed Hou10ni that solves harmonic wave equations with DG methods and curved elements. We found that both the effects of pollution and dispersion, which are very significant when a conventional finite element method is used, are limited [52]. However, bad conditioning is persisting and reliability of the method is not guaranteed when the coefficients vary considerably. In addition, the number of unknowns of the linear system is too big to hope to solve a realistic 3D problem. So it is important to develop approximation methods that require fewer degrees of freedom. Magique-3D wishes to invest heavily in the development of new approximation methods for harmonic wave equations. It is a difficult subject for which we want to develop different tasks, in collaboration with academic researchers with whom we are already working or have established contacts. Research directions that we would like to follow are the following.

First, we will continue our long-term collaboration with Prof. Rabia Djellouli. We want to continue to work on hybrid finite element methods that rely on basis functions composed of plane waves and polynomials. These methods have demonstrated good resistance to the phenomenon of numerical pollution [47], [48], but their capability of solving industrial problems has not been illustrated. This is certainly due to the absence of guideline for choosing the plane waves. We are thus currently working on the implementation of a methodology that makes the choice of plane waves automatic for a given simulation (fixed propagation domain, data source, etc.). This is up-front investigation and there is certainly a lot of remaining work before

being applied to geophysical imaging. But it gives the team the opportunity to test new ideas while remaining in contact with potential users of the methods.
Then we want to work with Prof. A. Bendali on developing methods of local integral equations which allow calculation of numerical fluxes on the edges of elements. One could then use these fluxes in a DG method for reconstructing the solution throughout the volume of calculation. This research is motivated by recent results which illustrate the difficulties of the existing methods which are not always able to approximate the propagating modes (plane waves) and the evanescent modes (polynomials) that may coexist, especially when one considers realistic applications. Integral equations are direct tools for computing fluxes and they are known for providing very good accuracy. They thus should help to improve the quality of approximation of DG methods which are fully flux-dependent. In addition, local integral equations would limit calculations at the interfaces, which would have the effect of limiting the number of unknowns generally high, especially for DG methods. Again, it is a matter of long-term research which success requires a significant amount of mathematical analysis, and also the development of non-trivial code.
To limit the effects of pollution and dispersion is not the only challenge that the team wants to tackle. Our experience alongside Total has made us aware of the difficulties in constructing meshes that are essential to achieve our simulations. There are several teams at Inria working on mesh generation and we are in contact with them, especially with Gamma3 (Paris-Rocquencourt Research Center). These teams develop meshes increasingly sophisticated to take account of the constraints imposed by realistic industrial benchmarks. But in our opinion, issues which are caused by the construction of meshes are not the only downside. Indeed, we have in mind to solve inverse problems and in this case it is necessary to mesh the domain at each iteration of Newton-type solver. It is therefore interesting to work on methods that either do not use mesh or rely on meshes which are very easy to construct. Regarding meshless methods, we have begun a collaboration with Prof. Djellouli which allowed us to propose a new approach called Mesh-based Frontier Free Formulation (MF3). The principle of this method is the use of fundamental solutions of Helmholtz equations as basic functions. One can then reduce the volumic variational formulation to a surfacic variational formulation which is close to an integral equation, but which does not require the calculation of singularities. The results are very promising and we hope to continue our study in the context of the application to geophysical imaging. An important step to validate this method will be particularly its extension to 3D because the results we have achieved so far are for 2D problems.
Keeping in mind the idea of limiting the difficulties of mesh, we want to study the method of virtual elements. This method attracts us because it relies on meshes that can be made of arbitrarily-shaped polygon and meshes should thus be fairly straightforward. Existing works on the subject have been mainly developed by the University of Pavia, in collaboration with Los Alamos National Laboratory [56], [57], [55], [53], [58]. None of them mentions the feasibility of the method for industrial applications and to our knowledge, there are no results on the method of virtual elements applied to the wave equations. First, we aim at applying the method described in [54] to the scalar Helmholtz equation and explore opportunities to use discontinuous elements within this framework. Then hp-adaptivity could be kept, which is particularly interesting for wave propagation in heterogeneous media.
DG methods are known to require a lot of unknowns that can exceed the limits accepted by the most advanced computers. This is particularly true for harmonic wave equations that require a large number of discretization points, even in the case of a conventional finite element method. We therefore wish to pursue a research activity that we have just started in collaboration with the project-team Nachos (Sophia-Antipolis Méditerranée Research Center). In order to reduce the number of degrees of freedom, we are interested in "hybrid mixed" Discontinuous Galerkin methods that provides a two-step procedure for solving the Helmholtz equations [69], [73], [72]. First, Lagrange multipliers are introduced to represent the flux of the numerical solution through the interface (edge or face) between two elements. The Lagrange multipliers are solution to a linear system which is constructed locally element by element. The number of degrees of freedom is then strongly reduced since for a standard DG method, there is a need of considering unknowns including volumetric values inside the element. And obviously, the gain is even more important when the order of the element is high. Next, the solution is reconstructed from the values of the multipliers and the cost of this step is negligible since it only requires inverting small-sized matrices. We have obtained promising results in the framework of the PhD

thesis of Marie Bonnasse-Gahot and we want to apply it to the simulation of complex phenomena such as the 3D viscoelastic wave propagation.
Obviously, the success of all these works depends on our ability to consider realistic applications such as wave propagation in the Earth. And in these cases, it is quite possible that even if we manage to develop accurate less expensive numerical methods, the solution of inverse problems will still be computationally intensive. It is thus absolutely necessary that we conduct our research by taking advantage of the latest advances in high-performance computing. We have already initiated discussions with the project team HIEPACS (Bordeaux Sud-Ouest research Center) to test the performance of the latest features of Mumps http://mumps.enseeiht.fr/, such as Low Rank Approximation or adaptation to hybrid CPU / GPU architectures and to Intel Xeon Phi, on realistic test cases. We are also in contact with the team Algorithm at Cerfacs (Toulouse) for the development of local integral equations solvers. These collaborations are essential for us and we believe that they will be decisive for the simulation of three-dimensional elasto-dynamic problems. However, our scientific contribution will be limited in this area because we are not experts in HPC.

3.4. Hybrid time discretizations of high-order

Most of the meshes we consider are composed of cells greatly varying in size. This can be due to the physical characteristics (propagation speed, topography, ...) which may require to refine the mesh locally, very unstructured meshes can also be the result of dysfunction of the mesher. For practical reasons which are essentially guided by the aim of reducing the number of matrix inversions, explicit schemes are generally privileged. However, they work under a stability condition, the so-called Courant Friedrichs Lewy (CFL) condition which forces the time step being proportional to the size of the smallest cell. Then, it is necessary to perform a huge number of iterations in time and in most of the cases because of a very few number of small cells. This implies to apply a very small time step on grids mainly composed of coarse cells and thus, there is a risk of creating numerical dispersion that should not exist. However, this drawback can be avoided by using low degree polynomial basis in space in the small meshes and high degree polynomials in the coarse meshes. By this way, it is possible to relax the CFL condition and in the same time, the dispersion effects are limited. Unfortunately, the cell-size variations are so important that this strategy is not sufficient. One solution could be to apply implicit and unconditionally stable schemes, which would obviously free us from the CFL constraint. Unfortunately, these schemes require inverting a linear system at each iteration and thus needs huge computational burden that can be prohibitive in 3D. Moreover, numerical dispersion may be increased. Then, as second solution is the use of local time stepping strategies for matching the time step to the different sizes of the mesh. There are several attempts [62], [59], [75], [71], [64] and Magique 3D has proposed a new time stepping method which allows us to adapt both the time step and the order of time approximation to the size of the cells. Nevertheless, despite a very good performance assessment in academic configurations, we have observed to our detriment that its implementation inside industrial codes is not obvious and in practice, improvements of the computational costs are disappointing, especially in a HPC framework. Indeed, the local time stepping algorithm may strongly affect the scalability of the code. Moreover, the complexity of the algorithm is increased when dealing with lossy media [67].

Recently, Dolean *et al* [63] have considered a novel approach consisting in applying hybrid schemes combining second order implicit schemes in the thin cells and second order explicit discretization in the coarse mesh. Their numerical results indicate that this method could be a good alternative but the numerical dispersion is still present. It would then be interesting to implement this idea with high-order time schemes to reduce the numerical dispersion. The recent arrival in the team of J. Chabassier should help us to address this problem since she has the expertise in constructing high-order implicit time scheme based on energy preserving Newmark schemes [60]. We propose that our work be organized around the two following tasks. The first one is the extension of these schemes to the case of lossy media because applying existing schemes when there is attenuation is not straightforward. This is a key issue because there is artificial attenuation when absorbing boundary conditions are introduced and if not, there are cases with natural attenuation like in visco-elastic media. The second one is the coupling of high-order implicit schemes with high-order explicit schemes. These two tasks can be first completed independently, but the ultimate goal is obviously to couple the schemes for lossy media. We will consider two strategies for the coupling. The first one will be based on the method

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

Another important issue is the load balancing that we can not avoid with parallel computing. In particular, we will have to choose one of these two alternatives: dedicate one part of processors to the implicit computations and the other one to explicit calculus or distribute the resolution with both schemes on all processors. A collaboration with experts in HPC is then mandatory since we are not expert in parallel computing. We will thus continue to collaborate with the team-projects Hiepacs and Runtime with whom we have a long-term experience of collaborations. The load-balancing leads then to the issue of mesh partitioning. Main mesh partitioners are very efficient for the coupling of different discretizations in space but to the best of our knowledge, the case of non-uniform time discretization has never been addressed. The study of meshes being out of the scopes of Magique-3D, we will collaborate with experts on mesh partitioning. We get already on to François Pellegrini who is the principal investigator of Scotch (http://www.labri.fr/perso/pelegrin/scotch) and permanent member of the team project Bacchus (Inria Bordeaux Sud Ouest Research Center).

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

SERENA Project-Team

3. Research Program

3.1. Multiphysics coupling

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

3.2. Structure-preserving discretizations and discrete element methods

We consequently design *numerical methods* for the devised model. Traditionally, we have worked in the context of finite element, finite volume, mixed finite element, and discontinuous Galerkin methods. Novel classes of schemes enable the use of general *polygonal* and *polyhedral meshes* with *nonmatching interfaces*, and we develop them in response to a high demand from our industrial partners (namely EDF, CEA, and IFP Energies Nouvelles). In the lowest-order case, our requirement is to derive *structure-preserving* methods, i.e., methods that mimic algebraically at the discrete level fundamental properties of the underlying PDEs, such as conservation principles and preservation of invariants. Here, the theoretical questions are closely linked to *differential geometry* and we apply them to the Navier–Stokes equations and to elasto-plasticity. In the higher-order case, we actively contribute to the development of hybrid high-order methods. We contribute to the numerical analysis in nonlinear cases (obstacle problem, Signorini conditions), we apply these methods to challenging problems from solid mechanics involving large deformations and plasticity, and we develop a comprehensive software implementing them. We believe that these methods belong to the future generation of numerical methods for industrial simulations; as a concrete example, the implementation of these methods in an industrial software of EDF has begun this year.

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

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

3.4. Reliability by a posteriori error control

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

3.5. Safe and correct programming

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

STEEP Project-Team

3. Research Program

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

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

In this context, to develop flexible integrated systemic models (upgradable, modular, ...) which are efficient, realistic and easy to use (for developers, modelers and end users) is a challenge in itself. What mathematical representations and what computational tools to use? Nowadays many tools are used: for example, cellular automata (e.g. in the LEAM model), agent models (e.g. URBANSIM [0]), system dynamics (e.g. World3), large systems of ordinary equations (e.g. equilibrium models such as TRANUS), and so on. Each of these tools has strengths and weaknesses. Is it necessary to invent other representations? What is the relevant level of modularity? How to get very modular models while keeping them very coherent and easy to calibrate? Is it preferable to use the same modeling tools for the whole system, or can we freely change the representation for each considered subsystem? How to easily and effectively manage different scales? (difficulty appearing in particular during the calibration process). How to get models which automatically adapt to the granularity of the data and which are always numerically stable? (this has also a direct link with the calibration processes and the propagation of uncertainties). How to develop models that can be calibrated with reasonable efforts, consistent with the (human and material) resources of the agencies and consulting firms that use them?

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

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

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

ecosystem service community has been using 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 has worked in particular on a land use/ land cover change (LUCC) modelling environments (Dinamica [0]) which belongs to the category of spatially explicit statistical models.

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

3.2. Model calibration and validation

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

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

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

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

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

3.3. Sensitivity analysis

[0] http://www.csr.ufmg.br/dinamica/

A sensitivity analysis (SA) consists, in a nutshell, in studying how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model inputs. It is complementary to an uncertainty analysis, which focuses on quantifying uncertainty in model output. SA's can be useful for several purposes, such as guiding model development and identifying the most influential model parameters and critical data items. Identifying influential model parameters may help in 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.

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

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

TONUS Team

3. Research Program

3.1. Kinetic models for plasmas

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

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

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

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

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

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

3.1.1. Gyrokinetic models: theory and approximation

In most phenomena where oscillations are present, we can establish a three-model hierarchy: (i) the model parameterized by the oscillation period, (ii) the limit model and (iii) the two-scale model, possibly with its corrector. In a context where one wishes to simulate such a phenomenon where the oscillation period is small and the oscillation amplitude is not small, it is important to have numerical methods based on an approximation of the Two-Scale model. If the oscillation period varies significantly over the domain of simulation, it is important to have numerical methods that approximate properly and effectively the model parameterized by the oscillation period and the Two-Scale model. Implementing Two-Scale Numerical Methods (for instance by Frénod et al. [22]) is based on the numerical approximation of the Two-Scale model. These are called of order 0. A Two-Scale Numerical Method is called of order 1 if it incorporates information from the corrector and from the equation of which this corrector is a solution. If the oscillation period varies between very small values and values of order 1, it is necessary to have new types of numerical schemes (Two-Scale Asymptotic Preserving Schemes of order 1 or TSAPS) that preserve the asymptotics between the model parameterized by the oscillation period and the Two-Scale model with its corrector. A first work in this direction has been initiated by Crouseilles et al. [21].

3.1.2. Semi-Lagrangian schemes

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

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

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

GYSELA is a 5D full gyrokinetic code based on a classical backward semi-Lagrangian scheme (BSL) [26] for the simulation of core turbulence that has been developed at CEA Cadarache in collaboration with our team [23].

More recently, we have started to apply the Semi-Lagrangian methods to more general kinetic equations. Indeed, most of the conservation laws of physics can be represented by a kinetic model with a small set of velocities and relaxation source terms [10]. Compressible fluids or MHD equations have such representations. Semi-Lagrangian methods then become a very appealing and efficient approach for solving these equations.

3.1.3. PIC methods

Historically PIC methods have been very popular for solving the Vlasov equations. They allow solving the equations in the phase space at a relatively low cost. The main disadvantage of this approach is that, due to its random aspect, it produces an important numerical noise that has to be controlled in some way, for instance by regularizations of the particles, or by divergence correction techniques in the Maxwell solver. We have a long-standing experience in PIC methods and we started implementing them in Selalib. An important aspect is to adapt the method to new multicore computers. See the work by Crestetto and Helluy [20].

3.2. Fluid and Reduced kinetic models for plasmas

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

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

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

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

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

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

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

Another way to reduce the model is to try to find an abstract kinetic representation with an as small as possible set of kinetic velocities. The kinetic approach has then only a mathematical meaning. It allows solving very efficiently many equations of physics [1].

3.2.1. Numerical schemes

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

In addition, this method, when written in a special set of variables, called the entropy variables, has nice properties concerning the entropy dissipation of the model. It opens the door to constructing numerical schemes with good conservation properties and no entropy dissipation, as already used for other systems of PDEs [27], [19], [25], [24].

3.2.2. Matrix-free Implicit schemes

In tokamaks, the reduced model generally involves many time scales. Among these time scales, many of then, associated to the fastest waves, are not relevant. In order to filter them out, it is necessary to adopt implicit solvers in time. When the reduced model is based on a kinetic interpretation, it is possible to construct implicit schemes that do not impose solving costly linear systems. In addition the resulting solver is stable even at very high CFL number [1].

3.3. Electromagnetic solvers

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

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

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

3.3.1. Coupling

The coupling of the Maxwell equations to the Vlasov solver requires some precautions. The most important one is to control the charge conservation errors, which are related to the divergence conditions on the electric and magnetic fields. We will generally use divergence correction tools for hyperbolic systems presented for instance in [17] (and the references therein).

3.3.2. Implicit solvers

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

BIOCORE Project-Team

3. Research Program

3.1. Mathematical and computational methods

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

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

In order to act on the system, which often is the purpose of our modeling approach, we then make use of two strong points of Control Theory: 1) the development of observers, that estimate the full internal state of the system from the measurements that we have, and 2) the design of a control law, that imposes to the system the behavior that we want to achieve, such as the regulation at a set point or optimization of its functioning. However, due to the peculiar structure and large uncertainties of our models, we need to develop specific methods. Since actual sensors can be quite costly or simply do not exist, a large part of the internal state often needs to be re-constructed from the measurements and one of the methods we developed consists in integrating the large uncertainties by assuming that some parameters or inputs belong to given intervals. We then developed robust observers that asymptotically estimate intervals for the state variables [81]. Using the directly measured variables and those that have been obtained through such, or other, observers, we then develop control methods that take advantage of the system structure (linked to competition or predation relationships between species in bioreactors or in the trophic networks created or modified by biological control).

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

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

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

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

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

CARMEN Project-Team

3. Research Program

3.1. Complex models for the propagation of cardiac action potentials

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

Several improvements of current models of the propagation of the action potential are being developed in the Carmen team, based on previous work [54] and on the data available at IHU LIRYC:

- Enrichment of the current monodomain and bidomain models [54], [63] by accounting for structural heterogeneities of the tissue at an intermediate scale. Here we focus on multiscale analysis techniques applied to the various high-resolution structural data available at the LIRYC.
- Coupling of the tissues from the different cardiac compartments and conduction systems. Here, we develop models that couple 1D, 2D and 3D phenomena described by reaction-diffusion PDEs.

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

We use these model codes for applied studies in two important areas of cardiac electrophysiology: atrial fibrillation [56] and sudden-cardiac-death (SCD) syndromes [7], [6] [59]. This work is performed in collaboration with several physiologists and clinicians both at IHU Liryc and abroad.

3.2. Simplified models and inverse problems

The medical and clinical exploration of the cardiac electric signals is based on accurate reconstruction of the patterns of propagation of the action potential. The correct detection of these complex patterns by non-invasive electrical imaging techniques has to be developed. This problem involves solving inverse problems that cannot be addressed with the more 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. Although widely studied recently, it remains a challenge for the scientific community. In many cases the quality of reconstructed electrical potential is not adequate. The methods used consist in solving the Laplace equation on the volume delimited by the body surface and the epicardial surface. Our aim is to

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

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

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

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

3.3. Numerical techniques

We want our numerical simulations to be efficient, accurate, and reliable with respect to the needs of the medical community. Based on previous work on solving the monodomain and bidomain equations [4], [5], [8], [1], we will focus on

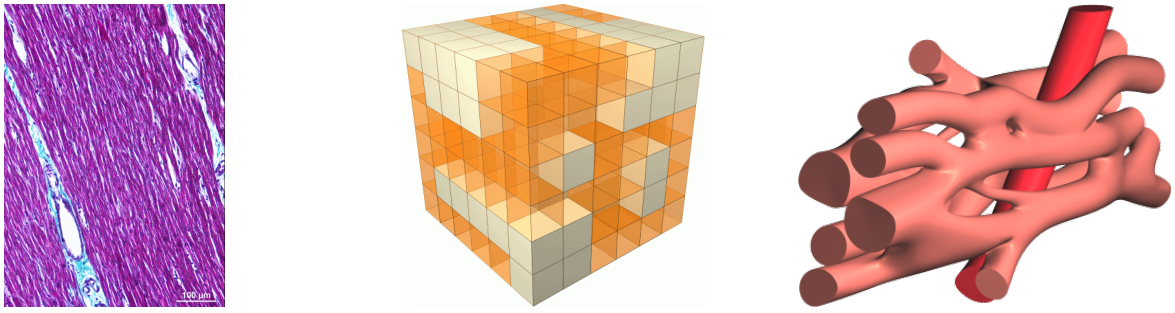
- High-order numerical techniques with respect to the variables with physiological meaning, like velocity, AP duration and restitution properties.
- Efficient, dedicated preconditioning techniques coupled with parallel computing.

Existing simulation tools used in our team rely, among others, on mixtures of explicit and implicit integration methods for ODEs, hybrid MPI-OpenMP parallellization, algebraic multigrid preconditioning, and Krylov solvers. New developments include high-order explicit integration methods and task-based dynamic parallelism.

3.4. Cardiac Electrophysiology at the Microscopic Scale

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

The literature on this type of model is still very limited [67]. Existing models are two-dimensional [60] or limited to idealized geometries, and use a linear (purely resistive) behaviour of the gap-juction channels that connect the cells. We propose a three-dimensional approach using realistic cellular geometry (figure 1), nonlinear gap-junction behaviour, and a numerical approach that can scale to hundreds of cells while maintaining a sub-micrometer spatial resolution (10 to 100 times smaller than the size of a cardiomyocyte) [52], [51], [50]. P-E. Bécue defended his PhD thesis on this topic in December 2018.



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

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

DRACULA Project-Team

3. Research Program

3.1. Mixed-effect models and statistical approaches

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

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

3.2. Development of a simulation platform

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

3.3. Mathematical and computational modeling

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

3.4. From hybrid dynamics to continuum mechanics

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

3.5. Structured partial differential equations

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

3.6. Delay differential equations

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

3.7. Multi-scale modeling of the immune response

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

We recently published a model of the CD8 T cell immune response characterizing differentiation stages, identified by biomarkers, able to predict the quantity of memory cells from early measurements ([32]). In parallel, we improved our multiscale model of the CD8 T cell immune response, by implementing a full differentiation scheme, from naïve to memory cells, based on a limited set of genes and transcription factors.

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

Our second task will be to update our multiscale model by first implementing the new differentiation scheme we identified ([32]), and second by embedding CD8 T cells with the GRN obtained in our first task (see above). This will lead to a multi-scale model incorporating description of the CD8 T cell immune response both at the molecular and the cellular levels (Simon Girel's PhD thesis, ongoing).

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

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

3.8. Dynamical network inference from single-cell data

Up to now, all of our multiscale models have incorporated a dynamical molecular network that was build "by hand" after a thorough review of the literature. It would be highly valuable to infer it directly from gene expression data. However, this remains very challenging from a methodological point of view. We started exploring an original solution for such inference by using the information contained within gene expression distributions. Such distributions can be acquired through novel techniques where gene expression levels are quantified at the single cell level. We propose to view the inference problem as a fitting procedure for a mechanistic gene network model that is inherently stochastic and takes not only protein, but also mRNA levels into account. This approach led to very encouraging results [34] and we will actively pursue in that direction, especially in the light of the foreseeable explosion of single cell data.

3.9. Leukemia modeling

Imatinib and other tyrosine kinase inhibitors (TKIs) have marked a revolution in the treatment of Chronic Myelogenous Leukemia (CML). Yet, most patients are not cured, and must take their treatment for life. Deeper mechanistic understanding could improve TKI combination therapies to better control the residual leukemic cell population. In a collaboration with the Hospital Lyon Sud and the University of Maryland, we have developed mathematical models that integrate CML and an autologous immune response ([29], [30] and [31]). These studies have lent theoretical support to the idea that the immune system plays a rôle in maintaining remission over long periods. Our mathematical model predicts that upon treatment discontinuation, the immune system can control the disease and prevent a relapse. There is however a possibility for relapse via a sneak-though mechanism [29]. Research in the next four years will focus in the Phase III PETALS trial. In the PETALS trial (https://clinicaltrials.gov/ct2/show/NCT02201459), the second generation TKI Nilotinib is combined with Peg-IFN, an interferon that is thought to enhance the immune response. We plan to: 1) Adapt the model to take into account the early dynamics (first three months). 2) Use a mixed-effect approach to analyse the effect of the combination, and find population and individual parameters related to treatment efficacy and immune system response. 3) Optimise long-term treatment strategies to reduce or cease treatment and make personalised predictions based on mixed-effect parameters, to minimise the long-term probability of relapse.

M3DISIM Project-Team

3. Research Program

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

Over the past decade, we have laid out the foundations of a multi-scale 3D model of the cardiac mechanical contraction responding to electrical activation. Several collaborations have been crucial in this enterprise, see below references. By integrating this formulation with adapted numerical methods, we are now able to represent the whole organ behavior in interaction with the blood during complete heart beats. This subject was our first achievement to combine a deep understanding of the underlying physics and physiology and our constant concern of proposing well-posed mathematical formulations and adequate numerical discretizations. In fact, we have shown that our model satisfies the essential thermo-mechanical laws, and in particular the energy balance, and proposed compatible numerical schemes that – in consequence – can be rigorously analyzed, see [6]. In the same spirit, we have formulated a poromechanical model adapted to the blood perfusion in the heart, hence precisely taking into account the large deformation of the mechanical medium, the fluid inertia and moving domain, and so that the energy balance between fluid and solid is fulfilled from the model construction to its discretization, see [7].

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

A major challenge in the context of biomechanical modeling – and more generally in modeling for life sciences – lies in using the large amount of data available on the system to circumvent the lack of absolute modeling ground truth, since every system considered is in fact patient-specific, with possibly non-standard conditions associated with a disease. We have already developed original strategies for solving this particular type of inverse problems by adopting the observer stand-point. The idea we proposed consists in incorporating to the classical discretization of the mechanical system an estimator filter that can use the data to improve the quality of the global approximation, and concurrently identify some uncertain parameters possibly related to a diseased state of the patient. Therefore, our strategy leads to a coupled model-data system solved similarly to a usual PDE-based model, with a computational cost directly comparable to classical Galerkin approximations. We have already worked on the formulation, the mathematical and numerical analysis of the resulting system – see [5] – and the demonstration of the capabilities of this approach in the context of identification of constitutive parameters for a heart model with real data, including medical imaging, see [3].

MAMBA Project-Team

3. Research Program

3.1. Introduction

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

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

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

Taking advantage of the last 4-year evaluation of MAMBA (September 2017), we have reorganized the presentation of our research program in three main methodological axes. Two main application axes are presented in the next Section. Evolving along their own logic in close interaction with the methodological axes, they are considered as application-driven research axes in themselves. The methodological research axes are the following.

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

Axis 2 is devoted to reaction equations and motion equations of agents in living systems. It aims at describing biological phenomena such as tumor growth, chemotaxis and wound healing.

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

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

Personnel

Pierre-Alexandre Bliman, Jean Clairambault, Marie Doumic, Benoît Perthame, Nastassia Pouradier Duteil, Philippe Robert

Project-team positioning

Population dynamics is a field with varied and wide applications, many of them being in the core of MAMBA interests - cancer, bacterial growth, protein aggregation. Their theoretical study also brings a qualitative understanding on the interplay between individual growth, propagation and reproduction in such populations. In the previous periods of evaluation, many results where obtained in the BANG team on the asymptotic and qualitative behavior of such structured population equations, see e.g. [126], [73], [94], [84]. Other Inria teams interested by this domain are Mycenae, Numed and Dracula, with which we are in close contacts. Among the leaders of the domain abroad, we can cite among others our colleagues Tom Banks (USA), Graeme Wake (New Zealand), Glenn Webb (USA), Jacek Banasiak (South Africa), Odo Diekmann (Netherlands), with whom we are also in regular contact. Most remarkably and recently, connections have also been made with probabilists working on Piecewise Deterministic Markov Processes (F. Malrieu at the university of Rennes, Jean Bertoin at the ETH in Zurich, Vincent Bansaye at Ecole Polytechnique, Julien Berestycki at Cambridge, Amaury Lambert at College de France, M. Hoffmann at Paris Dauphine), leading to a better understanding of the links between both types of results – see also the Methodological axis 3.

Scientific achievements

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

Time asymptotics for nucleation, growth and division equations

Following the many results obtained in the BANG team on the asymptotic and qualitative behavior of structured population equation, we put our effort on the investigation of limit cases, where the trend to a steady state or to a steady exponential growth described by the first eigenvector fails to happen. In [78], the case of equal mitosis (division into two equally-sized offspring) with linear growth rate was studied, and strangely enough, it appeared that the general relative entropy method could also be adapted to such a non-dissipative case. Many discussions and common workshops with probabilists, especially through the ANR project PIECE coordinated by F. Malrieu, have led both communities to work closer.

In [92], the case of constant fragmentation rate and linear growth rate has been investigated in a deterministic approach, whereas similar questions were simultaneously raised but in a stochastic process approach in [75].

We also enriched the models by taking into account a nucleation term, modeling the spontaneous formation of large polymers out of monomers [137]. We investigated the interplay between four processes: nucleation, polymerization, depolymerization and fragmentation.

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

Cell population dynamics and its control

One of the important incentives for such model design, source of many theoretical works, is the challenging question of drug-induced drug resistance in cancer cell populations, described in more detail below in the Applicative axis 1, Cancer. The adaptive dynamics setting used consists of phenotype-structured integro-differential [or reaction-diffusion, when phenotype instability is added under the form of a Laplacian] equations describing the dynamic behavior of different cell populations interacting in a Lotka-Volterra-like manner that represents common growth limitation due to scarcity of expansion space and nutrients. The phenotype structure allows us to analyse the evolution in phenotypic traits of the populations under study and its asymptotics for two populations [119], [116], [115], [117]. Space may be added as a complementary structure variable provided that something is known of the (Cartesian) geometry of the population [118], which is seldom the case.

Modelling, observation and identification of the spread of infectious diseases

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

Modelling Mendelian and non-Mendelian inheritances in density-dependent population dynamics
Classical strategies for controlling mosquitoes responsible of vector-borne disease are based on mechanical methods, such as elimination of oviposition sites; and chemical methods, such as insecticide spraying. Long term usage of the latter generates resistance [81], [103], transmitted to progeny according to Mendelian inheritance (in which each parent contributes randomly one of two possible alleles for a trait). New control strategies involve biological methods such as genetic control, which may either reduces mosquito population in a specific area or decreases the mosquito vector competence [61], [112], [144]. Among the latter, infection of wild populations by the bacterium *Wolbachia* appears promising (see also Applicative axis 2 below). Being maternally-transmitted, the latter obeys non-Mendelian inheritance law. Motivated by the effects of the (possibly unwanted) interaction of these two types of treatment, we initiated the study of modelling of Mendelian and non-Mendelian inheritances in density-dependent population dynamics.

Control of collective dynamics
The term *self-organization* is used to describe the emergence of complex organizational patterns from simple interaction rules in collective dynamics systems. Such systems are valuable tools to model various biological systems or opinion dynamics, whether it be the collective movement of animal groups, the organization of cells in an organism or the evolution of opinions in a large crowd. A special case of self-organization is given by *consensus*, i.e. the situation in which all agents' state variables converge. Another phenomenon is that of *clustering*, when the group is split into clusters that each converge to a different state. We have designed optimal control strategies to drive collective dynamics to consensus. In the case where consensus and clustering are situations to be avoided (for example in crowd dynamics), we designed control strategies to keep the system away from clustering.

Models of neural network
Mean field limits have been proposed by biophysicists in order to describe neural networks based on physiological models. The various resulting equations are called integrate-and-fire, time elapsed models, voltage-conductance models. Their specific nonlinearities and the blow-up phenomena make their originality which has led to develop specific mathematical analysis [129], followed by [124], [111], [130], [83]. This field also yields a beautiful illustration for the capacity of the team to combine and compare stochastic and PDE modelling (see Methodological axis 3), in [87].

Models of interacting particle systems
The organisation of biological tissues during development is accompanied by the formation of sharp borders between distinct cell populations. The maintenance of this cell segregation is key in adult tissue homeostatis, and its disruption can lead tumor cells to spread and form metastasis. This segregation is challenged during tissue growth and morphogenesis due to the high mobility of many cells that can lead to intermingling. Therefore, understanding the mechanisms involved in the generation and maintain of cell segregation is of tremendous importance in tissue morphogenesis, homeostasis, and in the development of various invasive diseases such as tumors. In this research axis, we aim to provide a mathematical framework which enables to quantitatively link the segregation and border sharpening ability of the tissue to these cell-cell interaction phenomena of interest [72]. As agent-based models do not enable precise mathematical analysis of their solutions due to the lack of theoretical results, we turn towards continuous -macroscopic- models and aim to provide a rigorous link between the different models [71].

Collaborations

- Nucleation, growth and fragmentation equations: **Juan Calvo**, university of Granada, came for two one-month visits, **Miguel Escobedo**, University of Bilbao (see also Methodological axis 3), **Pierre Gabriel**, University of Versailles-Saint Quentin, former B. Perthame and M. Doumic's Ph.D student, who now co-supervises Hugo Martin's Ph.D thesis. **Piotr Gwiązda**, Polish Academy of Sciences, Poland, **Emil Wiedemann**, University of Bonn, Germany, **Klemens Fellner**, university of Graz, Austria.
- Cell population dynamics and its control: **Tommaso Lorenzi**, former Mamba postdoc, now at the University of St. Andrews, Scotland, maintains a vivid collaboration with the Mamba team. He is in particular an external member of the HTE program MoGlImaging (see also Applicative axis 1). **Emmanuel Trélat**, Sorbonne Université professor, member of LJLL and of the CAGE Inria team, is the closest Mamba collaborator for optimal control. **Benedetto Piccoli**, Professor at Rutgers University (Camden, New Jersey), is collaborating on the analysis and control of collective dynamics.
- Mendelian inheritance and resistance in density-dependent population dynamics: **Pastor Pérez-Estigarribia, Christian Schaerer**, Universidad Nacional de Asunción, Paraguay.
- Neural networks: **Delphine Salort,** Professor Sorbonne Université, Laboratory for computations and quantification in biology, and **Patricia Reynaud**, University of Nice, **Maria Cáceres**, University of Granada.
- Models of interacting particle systems: **Pierre Degond**, Imperial College London, **MAPMO, Orléans**, **Ewelina Zatorska**, University College London, **Anais Khuong**, Francis Crick Institute

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

Personnel

Luis Almeida, Benoît Perthame, Diane Peurichard, Nastassia Pouradier Duteil.

Project-team positioning

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

Scientific achievements

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

Aggregation equation

In the mathematical study of collective behavior, an important class of models is given by the aggregation equation. In the presence of a non-smooth interaction potential, solutions of such systems may blow up in finite time. To overcome this difficulty, we have defined weak measure-valued solutions in the sense of duality and its equivalence with gradient flows and entropy solutions in one dimension [109]. The extension to higher dimensions has been studied in [86]. An interesting consequence of this approach is the possibility to use the traditional finite volume approach to design numerical schemes able to capture the good behavior of such weak measure-valued solutions [102], [108].

Identification of the mechanisms of single cell motion.

In this research axis, we aim to study the mechanisms of single cell adhesion-based and adhesion free motion. This work is done in the frame of the recently created associated team MaMoCeMa (see Section 9) with the WPI, Vienna. In a first direction [140] with N. Sfakianakis (Heidelberg University), we extended the live-cell motility Filament Based Lamellipodium Model to incorporate the forces exerted on the lamellipodium of the cells due to cell-cell collision and cadherin induced cell-cell adhesion. We took into account the nature of these forces via physical and biological constraints and modelling assumptions. We investigated the effect these new components had in the migration and morphology of the cells through particular experiments. We exhibit moreover the similarities between our simulated cells and HeLa cancer cells.

In a second work done in collaboration with the group of biologist at IST (led by **Michael Sixt** Austria), we developed and analyzed a two-dimensional mathematical model for cells migrating without adhesion capabilities [110]. Cells are represented by their cortex, which is modelled as an elastic curve, subject to an internal pressure force. Net polymerization or depolymerization in the cortex is modelled via local addition or removal of material, driving a cortical flow. The model takes the form of a fully nonlinear degenerate parabolic system. An existence analysis is carried out by adapting ideas from the theory of gradient flows. Numerical simulations show that these simple rules can account for the behavior observed in experiments, suggesting a possible mechanical mechanism for adhesion-independent motility.

Free boundary problems for tumor growth.
Fluid dynamic equations are now commonly used to describe tumor growth with two main classes of models: those which describe tumor growth through the dynamics of the density of tumoral cells subjected to a mechanical stress; those describing the tumor through the dynamics of its geometrical domain thanks to a Hele-Shaw-type free boundary model. The first link between these two classes of models has been rigorously obtained thanks to an incompressible limit in [128] for a simple model. This result has motivated the use of another strategy based on viscosity solutions, leading to similar results, in [113].

Since more realistic systems are used in the analysis of medical images, we have extended these studies to include active motion of cells in [127], viscosity in [132] and proved regularity results in [120]. The limiting Hele-Shaw free boundary model has been used to describe mathematically the invasion capacity of a tumour by looking for travelling wave solutions, in [131], see also Methodological axis 3. It is a fundamental but difficult issue to explain rigorously the emergence of instabilities in the direction transversal to the wave propagation. For a simplified model, a complete explanation is obtained in [114].

Two-way coupling of diffusion and growth.
We are currently developing a mathematical framework for diffusion equations on time-evolving manifolds, where the evolution of the manifold is a function of the distribution of the diffusing quantity. The need for such a framework takes it roots in developmental biology. Indeed, the growth of an organism is triggered by signaling molecules called morphogens that diffuse in the organism during its development. Meanwhile, the diffusion of the morphogens is itself affected by the changes in shape and size of the organism. In other words, there is a complete coupling between the diffusion of the morphogens and the evolution of the shapes. In addition to the elaboration of this theoretical framework, we also collaborate with a team of developmental biologists from Rutgers University (Camden, New Jersey) to develop a model for the diffusion of Gurken during the oogenesis of *Drosophila*.

Collaborations

- Shanghai Jiao Tong University, joint publications with Min Tang on bacterial models for chemotaxis and free boundary problems for tumor growth.
- Imperial College London, joint works with José Antonio Carrillo on aggregation equation.
- University of Maryland at College Park, UCLA, Univ. of Chicago, Univ. Autónoma de Madrid, Univ. of St. Andrews (Scotland), joint works on mathematics of tumor growth models.
- Joint work with Francesco Rossi (Università di Padova, Italy) and Benedetto Piccoli (Rutgers University, Camden, New Jersey, USA) on Developmental PDEs.
- Cooperation with Shugo Yasuda (University of Hyogo, Kobe, Japan) and Vincent Calvez (EPI Dracula) on the subject of bacterial motion.

- Cooperation with Nathalie Ferrand (INSERM), Michèle Sabbah (INSERM) and Guillaume Vidal (Centre de Recherche Paul Pascal, Bordeaux) on cell aggregation by chemotaxis.
- Nicolas Vauchelet, Université Paris 13

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

Personnel

Marie Doumic, Dirk Drasdo.

Project-team positioning

Mamba developed and addressed model and parameter identification methods and strategies in a number of mathematical and computational model applications including growth and fragmentation processes emerging in bacterial growth and protein misfolding, in liver regeneration [97], TRAIL treatment of HeLa cells [74], growth of multicellular spheroids [107], blood detoxification after drug-induced liver damage [139], [101].

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

Scientific achievements

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

Estimation methods for growing and dividing populations

In this domain, all originated in two papers in collaboration with J.P. Zubelli in 2007 [133], [96], whose central idea was to used the asymptotic steady distribution of the individuals to estimate the division rate. A series of papers improved and extended these first results while keeping the deterministic viewpoint, lastly [78]. The last developments now tackle the still more involved problem of estimating not only the division rate but also the fragmentation kernel (i.e., how the sizes of the offspring are related to the size of the dividing individual) [13]. In parallel, in a long-run collaboration with statisticians, we studied the Piecewise Deterministic Markov Process (PDMP) underlying the equation, and estimated the division rate directly on sample observations of the process, thus making a bridge between the PDE and the PDMP approach in [95], a work which inspired also very recently other groups in statistics and probability [75], [105] and was the basis for Adélaïde Olivier's Ph.D thesis [122], [106] and of some of her more recent works [123] (see also axis 5).

Data assimilation and stochastic modeling for protein aggregation

Estimating reaction rates and size distributions of protein polymers is an important step for understanding the mechanisms of protein misfolding and aggregation (see also axis 5). In [63], we settled a framework problem when the experimental measurements consist in the time-dynamics of a moment of the population.

To model the intrinsic variability among experimental curves in aggregation kinetics - an important and poorly understood phenomenon - Sarah Eugène's Ph.D, co-supervised by P. Robert [99], was devoted to the stochastic modeling and analysis of protein aggregation, compared both with the deterministic approach traditionally developed in Mamba [137] and with experiments.

Statistical methods decide on subsequently validated mechanism of ammonia detoxification

To identify the mechanisms involved in ammonia detoxification [101], 8 candidate models representing the combination of three possible mechanisms were developed (axis 5). First, the ability of each model to capture the experimental data was assessed by statistically testing the null hypothesis that the data have been generated by the model, leading to exclusion of one of the 8 models. The 7 remaining models were compared among each other by the likelihood ratio. The by far best models were those containing a particular ammonia sink mechanism, later validated experimentally (axis 5). For each of the statistical tests, the corresponding test statistics has been calculated empirically and turned out to be not chi2-distributed in opposition to the usual assumption stressing the importance of calculating the empirical distribution, especially when some parameters are unidentifiable. This year the ammonia detoxification mechanisms have been integrated in a spatial-temporal agent-based model of a liver lobule (the smallest repetitive anatomical unit of liver) and studied for normal and fibrotic liver.

Collaborations

- **Marc Hoffmann,** Université Paris-Dauphine, for the statistical approach to growth and division processes [95], **M. Escobedo**, Bilbao and **M. Tournus**, Marseille, for the deterministic approach.
- **Tom Banks**, North Carolina State University, and **Philippe Moireau,** Inria M3DISIM, for the inverse problem and data assimilation aspects [70], [62]
- **Jan G. Hengstler**, IfADo, Dortmund, Germany

MONC Project-Team

3. Research Program

3.1. Introduction

We are working in the context of data-driven medicine against cancer. We aim at coupling mathematical models with data to address relevant challenges for biologists and clinicians in order for instance to improve our understanding in cancer biology and pharmacology, assist the development of novel therapeutic approaches or develop personalized decision-helping tools for monitoring the disease and evaluating therapies.

More precisely, our research on mathematical oncology is three-fold:

- Axis 1: Tumor modeling for patient-specific simulations: *Clinical monitoring. Numerical markers from imaging data. Radiomics*.
- Axis 2: Bio-physical modeling for personalized therapies: *Electroporation from cells to tissue. Radiotherapy*.
- Axis 3: Quantitative cancer modeling for biological and preclinical studies: *Biological mechanisms. Metastatic dissemination. Pharmacometrics*.

In the first axis, we aim at producing patient-specific simulations of the growth of a tumor or its response to treatment starting from a series of images. We hope to be able to offer a valuable insight on the disease to the clinicians in order to improve the decision process. This would be particularly useful in the cases of relapses or for metastatic diseases.

The second axis aims at modeling biophysical therapies like radiotherapies, but also thermo-ablations, radio-frequency ablations or electroporation that play a crucial role in the case of a relapse or for a metastatic disease, which is precisely the clinical context where the techniques of axis 1 will be applied.

The third axis, even if not directly linked to clinical perspectives, is essential since it is a way to better understand and model the biological reality of cancer growth and the (possibly complex) effects of therapeutic intervention. Modeling in this case also helps to interpret the experimental results and improve the accuracy of the models used in Axis 1. Technically speaking, some of the computing tools are similar to those of Axis 1.

3.2. Axis 1: Tumor modeling for patient-specific simulations

The gold standard treatment for most cancers is surgery. In the case where total resection of the tumor is possible, the patient often benefits from an adjuvant therapy (radiotherapy, chemotherapy, targeted therapy or a combination of them) in order to eliminate the potentially remaining cells that may not be visible. In this case personalized modeling of tumor growth is useless and statistical modeling will be able to quantify the risk of relapse, the mean progression-free survival time...However if total resection is not possible or if metastases emerge from distant sites, clinicians will try to control the disease for as long as possible. A wide set of tools are available. Clinicians may treat the disease by physical interventions (radiofrequency ablation, cryoablation, radiotherapy, electroporation, focalized ultrasound,...) or chemical agents (chemotherapies, targeted therapies, antiangiogenic drugs, immunotherapies, hormonotherapies). One can also decide to monitor the patient without any treatment (this is the case for slowly growing tumors like some metastases to the lung, some lymphomas or for some low grade glioma). A reliable patient-specific model of tumor evolution with or without therapy may have different uses:

- Case without treatment: the evaluation of the growth of the tumor would offer a useful indication for the time at which the tumor will reach a critical size. For example, radiofrequency ablation of pulmonary lesion is very efficient as long as the diameter of the lesion is smaller than 3 cm. Thus, the prediction can help the clinician plan the intervention. For slowly growing tumors, quantitative modeling can also help to decide at what time interval the patient has to undergo a CT-scan. CT-scans

are irradiative exams and there is a challenge for decreasing their occurrence for each patient. It has also an economical impact. And if the disease evolution starts to differ from the prediction, this might mean that some events have occurred at the biological level. For instance, it could be the rise of an aggressive phenotype or cells that leave a dormancy state. This kind of events cannot be predicted, but some mismatch with respect to the prediction can be an indirect proof of their existence. It could be an indication for the clinician to start a treatment.

- Case with treatment: a model can help to understand and to quantify the final outcome of a treatment using the early response. It can help for a redefinition of the treatment planning. Modeling can also help to anticipate the relapse by analyzing some functional aspects of the tumor. Again, a deviation with respect to reference curves can mean a lack of efficiency of the therapy or a relapse. Moreover, for a long time, the response to a treatment has been quantified by the RECIST criteria which consists in (roughly speaking) measuring the diameters of the largest tumor of the patient, as it is seen on a CT-scan. This criteria is still widely used and was quite efficient for chemotherapies and radiotherapies that induce a decrease of the size of the lesion. However, with the systematic use of targeted therapies and anti-angiogenic drugs that modify the physiology of the tumor, the size may remain unchanged even if the drug is efficient and deeply modifies the tumor behavior. One better way to estimate this effect could be to use functional imaging (Pet-scan, perfusion or diffusion MRI, ...), a model can then be used to exploit the data and to understand in what extent the therapy is efficient.
- Optimization: currently, we do not believe that we can optimize a particular treatment in terms of distribution of doses, number, planning with the model that we will develop in a medium term perspective. But it is an aspect that we keep in mind on a long term one.

The scientific challenge is therefore as follows: knowing the history of the patient, the nature of the primitive tumor, its histopathology, knowing the treatments that patients have undergone, knowing some biological facts on the tumor and having a sequence of images (CT-scan, MRI, PET or a mix of them), are we able to provide a numerical simulation of the extension of the tumor and of its metabolism that fits as best as possible with the data (CT-scans or functional data) and that is predictive in order to address the clinical cases described above?

Our approach relies on the elaboration of PDE models and their parametrization with the image by coupling deterministic and stochastic methods. The PDE models rely on the description of the dynamics of cell populations. The number of populations depends on the pathology. For example, for glioblastoma, one needs to use proliferative cells, invasive cells, quiescent cells as well as necrotic tissues to be able to reproduce realistic behaviors of the disease. In order to describe the relapse for hepatic metastases of gastro-intestinal stromal tumor (gist), one needs three cell populations: proliferative cells, healthy tissue and necrotic tissue.

The law of proliferation is often coupled with a model for the angiogenesis. However such models of angiogenesis involve too many non measurable parameters to be used with real clinical data and therefore one has to use simplified or even simplistic versions. The law of proliferation often mimics the existence of an hypoxia threshold, it consists of an ODE. or a PDE that describes the evolution of the growth rate as a combination of sigmoid functions of nutrients or roughly speaking oxygen concentration. Usually, several laws are available for a given pathology since at this level, there are no quantitative argument to choose a particular one.

The velocity of the tumor growth differs depending on the nature of the tumor. For metastases, we will derive the velocity thanks to Darcy's law in order to express that the extension of the tumor is basically due to the increase of volume. This gives a sharp interface between the metastasis and the surrounding healthy tissues, as observed by anatomopathologists. For primitive tumors like glioma or lung cancer, we use reaction-diffusion equations in order to describe the invasive aspects of such primitive tumors.

The modeling of the drugs depends on the nature of the drug: for chemotherapies, a death term can be added into the equations of the population of cells, while antiangiogenic drugs have to be introduced in a angiogenic model. Resistance to treatment can be described either by several populations of cells or with non-constant growth or death rates. As said before, it is still currently difficult to model the changes of phenotype

or mutations, we therefore propose to investigate this kind of phenomena by looking at deviations of the numerical simulations compared to the medical observations.

The calibration of the model is achieved by using a series (at least 2) of images of the same patient and by minimizing a cost function. The cost function contains at least the difference between the volume of the tumor that is measured on the images with the computed one. It also contains elements on the geometry, on the necrosis and any information that can be obtained through the medical images. We will pay special attention to functional imaging (PET, perfusion and diffusion MRI). The inverse problem is solved using a gradient method coupled with some Monte-Carlo type algorithm. If a large number of similar cases is available, one can imagine to use statistical algorithms like random forests to use some non quantitative data like the gender, the age, the origin of the primitive tumor...for example for choosing the model for the growth rate for a patient using this population knowledge (and then to fully adapt the model to the patient by calibrating this particular model on patient data) or for having a better initial estimation of the modeling parameters. We have obtained several preliminary results concerning lung metastases including treatments and for metastases to the liver.

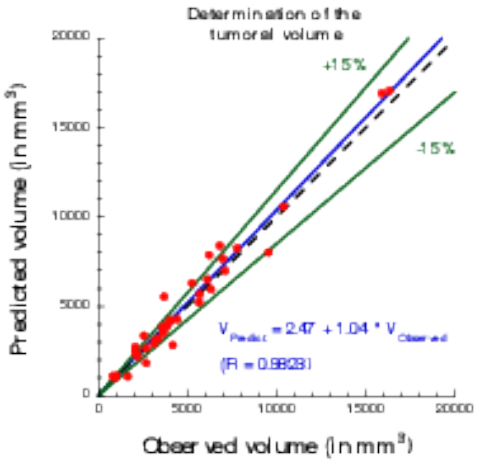


Figure 4. Plot showing the accuracy of our prediction on meningioma volume. Each point corresponds to a patient whose two first exams were used to calibrate our model. A patient-specific prediction was made with this calibrated model and compared with the actual volume as measured on a third time by clinicians. A perfect prediction would be on the black dashed line. Medical data was obtained from Prof. Loiseau, CHU Pellegrin.

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

In this axis, we investigate locoregional therapies such as radiotherapy, irreversible electroporation. Electroporation consists in increasing the membrane permeability of cells by the delivery of high voltage pulses. This non-thermal phenomenon can be transient (reversible) or irreversible (IRE). IRE or electro-chemotherapy – which is a combination of reversible electroporation with a cytotoxic drug – are essential tools for the treatment of a metastatic disease. Numerical modeling of these therapies is a clear scientific challenge. Clinical applications of the modeling are the main target, which thus drives the scientific approach, even though theoretical studies in order to improve the knowledge of the biological phenomena, in particular for electroporation, should also be addressed. However, this subject is quite wide and we focus on two particular approaches: some aspects of radiotherapies and electro-chemotherapy. This choice is motivated partly by pragmatic reasons: we

already have collaborations with physicians on these therapies. Other treatments could be probably treated with the same approach, but we do not plan to work on this subject on a medium term.

- Radiotherapy (RT) is a common therapy for cancer. Typically, using a CT scan of the patient with the structures of interest (tumor, organs at risk) delineated, the clinicians optimize the dose delivery to treat the tumor while preserving healthy tissues. The RT is then delivered every day using low resolution scans (CBCT) to position the beams. Under treatment the patient may lose weight and the tumor shrinks. These changes may affect the propagation of the beams and subsequently change the dose that is effectively delivered. It could be harmful for the patient especially if sensitive organs are concerned. In such cases, a replanification of the RT could be done to adjust the therapeutical protocol. Unfortunately, this process takes too much time to be performed routinely. The challenges faced by clinicians are numerous, we focus on two of them:
  - *Detecting the need of replanification:* we are using the positioning scans to evaluate the movement and deformation of the various structures of interest. Thus we can detect whether or not a structure has moved out of the safe margins (fixed by clinicians) and thus if a replanification may be necessary. In a retrospective study, our work can also be used to determine RT margins when there are no standard ones. A collaboration with the RT department of Institut Bergonié is underway on the treatment of retroperitoneal sarcoma and ENT tumors (head and neck cancers). A retrospective study was performed on 11 patients with retro-peritoneal sarcoma. The results have shown that the safety margins (on the RT) that clinicians are currently using are probably not large enough. The tool used in this study was developed by an engineer funded by Inria (Cynthia Périer, ADT Sesar). We used well validated methods from a level-set approach and segmentation / registration methods. The originality and difficulty lie in the fact that we are dealing with real data in a clinical setup. Clinicians have currently no way to perform complex measurements with their clinical tools. This prevents them from investigating the replanification. Our work and the tools developed pave the way for easier studies on evaluation of RT plans in collaboration with Institut Bergonié. *There was no modeling involved in this work that arose during discussions with our collaborators.* The main purpose of the team is to have meaningful outcomes of our research for clinicians, sometimes it implies leaving a bit our area of expertise.
  - *Evaluating RT efficacy and finding correlation between the radiological responses and the clinical outcome:* our goal is to help doctors to identify correlation between the response to RT (as seen on images) and the longer term clinical outcome of the patient. Typically, we aim at helping them to decide when to plan the next exam after the RT. For patients whose response has been linked to worse prognosis, this exam would have to be planned earlier. This is the subject of collaborations with Institut Bergonié and CHU Bordeaux on different cancers (head and neck, pancreas). The response is evaluated from image markers (*e.g.* using texture information) or with a mathematical model developed in Axis 1. The other challenges are either out of reach or not in the domain of expertise of the team. Yet our works may tackle some important issues for adaptive radiotherapy.
- Both IRE and electrochemotherapy are anticancerous treatments based on the same phenomenon: the electroporation of cell membranes. This phenomenon is known for a few decades but it is still not well understood, therefore our interest is two fold:
  1. We want to use mathematical models in order to better understand the biological behavior and the effect of the treatment. We work in tight collaboration with biologists and bioeletromagneticians to derive precise models of cell and tissue electroporation, in the continuity of the research program of the Inria team-project MC2. These studies lead to complex non-linear mathematical models involving some parameters (as less as possible). Numerical methods to compute precisely such models and the calibration of the parameters with the experimental data are then addressed. Tight collaborations with the Vectorology and Anticancerous Therapies (VAT) of IGR at Villejuif, Laboratoire Ampère of Ecole

Centrale Lyon and the Karlsruhe Institute of technology will continue, and we aim at developing new collaborations with Institute of Pharmacology and Structural Biology (IPBS) of Toulouse and the Laboratory of Molecular Pathology and Experimental Oncology (LMPEO) at CNR Rome, in order to understand differences of the electroporation of healthy cells and cancer cells in spheroids and tissues.

2. This basic research aims at providing new understanding of electroporation, however it is necessary to address, particular questions raised by radio-oncologists that apply such treatments. One crucial question is "What pulse or what train of pulses should I apply to electroporate the tumor if the electrodes are located as given by the medical images"? Even if the real-time optimization of the placement of the electrodes for deep tumors may seem quite utopian since the clinicians face too many medical constraints that cannot be taken into account (like the position of some organs, arteries, nerves...), one can expect to produce real-time information of the validity of the placement done by the clinician. Indeed, once the placement is performed by the radiologists, medical images are usually used to visualize the localization of the electrodes. Using these medical data, a crucial goal is to provide a tool in order to compute in real-time and visualize the electric field and the electroporated region directly on theses medical images, to give the doctors a precise knowledge of the region affected by the electric field. In the long run, this research will benefit from the knowledge of the theoretical electroporation modeling, but it seems important to use the current knowledge of tissue electroporation – even quite rough –, in order to rapidly address the specific difficulty of such a goal (real-time computing of non-linear model, image segmentation and visualization). Tight collaborations with CHU Pellegrin at Bordeaux, and CHU J. Verdier at Bondy are crucial.

- Radiofrequency ablation. In a collaboration with Hopital Haut Leveque, CHU Bordeaux we are trying to determine the efficacy and risk of relapse of hepatocellular carcinoma treated by radiofrequency ablation. For this matter we are using geometrical measurements on images (margins of the RFA, distance to the boundary of the organ) as well as texture information to statistically evaluate the clinical outcome of patients.
- Intensity focused ultrasound. In collaboration with Utrecht Medical center, we aim at tackling several challenges in clinical applications of IFU: target tracking, dose delivery...

3.4. Axis 3: Quantitative cancer modeling for biological and preclinical studies

With the emergence and improvement of a plethora of experimental techniques, the molecular, cellular and tissue biology has operated a shift toward a more quantitative science, in particular in the domain of cancer biology. These quantitative assays generate a large amount of data that call for theoretical formalism in order to better understand and predict the complex phenomena involved. Indeed, due to the huge complexity underlying the development of a cancer disease that involves multiple scales (from the genetic, intra-cellular scale to the scale of the whole organism), and a large number of interacting physiological processes (see the so-called "hallmarks of cancer"), several questions are not fully understood. Among these, we want to focus on the most clinically relevant ones, such as the general laws governing tumor growth and the development of metastases (secondary tumors, responsible of 90% of the deaths from a solid cancer). In this context, it is thus challenging to exploit the diversity of the data available in experimental settings (such as *in vitro* tumor spheroids or *in vivo* mice experiments) in order to improve our understanding of the disease and its dynamics, which in turn lead to validation, refinement and better tuning of the macroscopic models used in the axes 1 and 2 for clinical applications.

In recent years, several new findings challenged the classical vision of the metastatic development biology, in particular by the discovery of organism-scale phenomena that are amenable to a dynamical description in terms of mathematical models based on differential equations. These include the angiogenesis-mediated distant inhibition of secondary tumors by a primary tumor the pre-metastatic niche or the self-seeding phenomenon Building a general, cancer type specific, comprehensive theory that would integrate these dynamical processes

remains an open challenge. On the therapeutic side, recent studies demonstrated that some drugs (such as the Sunitinib), while having a positive effect on the primary tumor (reduction of the growth), could *accelerate* the growth of the metastases. Moreover, this effect was found to be scheduling-dependent. Designing better ways to use this drug in order to control these phenomena is another challenge. In the context of combination therapies, the question of the *sequence* of administration between the two drugs is also particularly relevant.

One of the technical challenge that we need to overcome when dealing with biological data is the presence of potentially very large inter-animal (or inter-individual) variability.

Starting from the available multi-modal data and relevant biological or therapeutic questions, our purpose is to develop adapted mathematical models (*i.e.* identifiable from the data) that recapitulate the existing knowledge and reduce it to its more fundamental components, with two main purposes:

1. to generate quantitative and empirically testable predictions that allow to assess biological hypotheses or
2. to investigate the therapeutic management of the disease and assist preclinical studies of anti-cancerous drug development.

We believe that the feedback loop between theoretical modeling and experimental studies can help to generate new knowledge and improve our predictive abilities for clinical diagnosis, prognosis, and therapeutic decision. Let us note that the first point is in direct link with the axes 1 and 2 of the team since it allows us to experimentally validate the models at the biological scale (*in vitro* and *in vivo* experiments) for further clinical applications.

More precisely, we first base ourselves on a thorough exploration of the biological literature of the biological phenomena we want to model: growth of tumor spheroids, *in vivo* tumor growth in mice, initiation and development of the metastases, effect of anti-cancerous drugs. Then we investigate, using basic statistical tools, the data we dispose, which can range from: spatial distribution of heterogeneous cell population within tumor spheroids, expression of cell makers (such as green fluorescent protein for cancer cells or specific antibodies for other cell types), bioluminescence, direct volume measurement or even intra-vital images obtained with specific imaging devices. According to the data type, we further build dedicated mathematical models that are based either on PDEs (when spatial data is available, or when time evolution of a structured density can be inferred from the data, for instance for a population of tumors) or ODEs (for scalar longitudinal data). These models are confronted to the data by two principal means:

1. when possible, experimental assays can give a direct measurement of some parameters (such as the proliferation rate or the migration speed) or
2. statistical tools to infer the parameters from observables of the model.

This last point is of particular relevance to tackle the problem of the large inter-animal variability and we use adapted statistical tools such as the mixed-effects modeling framework.

Once the models are shown able to describe the data and are properly calibrated, we use them to test or simulate biological hypotheses. Based on our simulations, we then aim at proposing to our biological collaborators new experiments to confirm or infirm newly generated hypotheses, or to test different administration protocols of the drugs. For instance, in a collaboration with the team of the professor Andreas Bikfalvi (Laboratoire de l'Angiogénèse et du Micro-environnement des Cancers, Inserm, Bordeaux), based on confrontation of a mathematical model to multi-modal biological data (total number of cells in the primary and distant sites and MRI), we could demonstrate that the classical view of metastatic dissemination and development (one metastasis is born from one cell) was probably inaccurate, in mice grafted with metastatic kidney tumors. We then proposed that metastatic germs could merge or attract circulating cells. Experiments involving cells tagged with two different colors are currently performed in order to confirm or infirm this hypothesis.

Eventually, we use the large amount of temporal data generated in preclinical experiments for the effect of anti-cancerous drugs in order to design and validate mathematical formalisms translating the biological mechanisms of action of these drugs for application to clinical cases, in direct connection with the axis 1. We have a special focus on targeted therapies (designed to specifically attack the cancer cells while sparing the

healthy tissue) such as the Sunitinib. This drug is indeed indicated as a first line treatment for metastatic renal cancer and we plan to conduct a translational study coupled between A. Bikfalvi's laboratory and medical doctors, F. Cornelis (radiologist) and A. Ravaud (head of the medical oncology department).

NUMED Project-Team

3. Research Program

3.1. Design of complex models

3.1.1. Project team positioning

The originality of our work is the quantitative description of phenomena accounting for several time and spatial scales. Here, propagation has to be understood in a broad sense. This includes propagation of invasive species, chemotactic waves of bacteira, evoluation of age structures populations ... Our main objectives are the quantitative calculation of macroscopic quantities as the rate of propagation, and microscopic distributions at the edge and the back of the front. These are essential features of propagation which are intimately linked in the long time dynamics.

3.1.2. Recent results

- Population models.

  H. Leman works at the interface between mathematics and biology, thanks to probabilist and determinist studies of models of populations. More precisely, she studies and develops probabilistic models, called agent models that described the population at an individual level. Each individual is characterized by one or more phenotypic traits and by its position, which may influence at the same time its ecological behavior and its motion. From a biological point of view these models are particularly interesting since they allow to include a large variety of interactions between individuals. These processes may also be studied in details to obtain theoretical results which may be simulated thanks to exact algorithms. To get quantitative results H. Leman uses changes of scales in space and time (large population, rare mutations, long time), following various biological assumptions.

  In a first study , H. Leman tries to understand the interactions between sexual preference mechanisms and evolutive forces inside spatially structured populations. Recently she got interesting in the description of necessary conditions to facilitate the emergence of such preferences by individuals.

  As a second example, H. Leman is also interested in the modeling and study of cooperative bacterias and tries to understand the impact of spatial structures in the eco - evolutions of these bacterias. Space seems to be an essential factor to facilitate the emergence of cooperation between bacteries.
- Inviscid limit of Navier Stokes equations.

  The question of the behavior of solutions of Navier Stokes equations in a bounded domain as the viscosity goes to 0 is a classical and highly difficult open question in Fluid Mechanics. A small boundary layer, called Prandtl layer, appears near the boundary, which turns out to be unstable if the viscosity is small enough. The stability analysis of this boundary layer is highly technical and remained open since the first formal analysis in the 1940's by physicists like Orr, Sommerfeld, Tollmien, Schlichting or Lin. E. Grenier recently made a complete mathematical analysis of this spectral problem, in collaboration with T. Nguyen and Y. Guo. We rigorously proved that any shear layer is spectrally and linearly unstable if the viscosity is small enough, which is the first mathematical result in that field. We also get some preliminary nonlinear results. A book on this subject is in preparation, already accepted by Springer.
- Numerical analysis of complex fluids: the example of avalanches.

  This deals with the development of numerical schemes for viscoplastic materials (namely with Bingham or Herschell-Bulkley laws). Recently, with other colleagues, Paul Vigneaux finished the design of the first 2D well-balanced finite volume scheme for a shallow viscoplastic model. It is illustrated on the famous Taconnaz avalanche path in the Mont-Blanc, Chamonix, in the case of dense snow avalanches. The scheme deals with general Digital Elevation Model (DEM) topographies, wet/dry fronts and is designed to compute precisely the stopping state of avalanches, a crucial point of viscoplastic flows which are able to rigidify [21].

3.1.3. Collaborations

- Ecology: Orsay (C. Coron), Toulouse (IMT, M. Costa), MNHM Paris (V. Llaurens), LISC Paris (C. Smadi), ENS Paris (R. Ferrière, E. Abs), CIMAT (Mexique, J. C. P. Millan).
- Inviscid limit of Navier Stokes equations: Brown University (Y. Guo, B. Pausader), Penn State University (T. Nguyen), Orsay University (F. Rousset).
- Numerical analysis of complex fluids: Enrique D. Fernandez - Nieto (Univ. de Sevilla, Spain), Jose Maria Gallardo (Univ. de Malaga, Spain).

3.2. Parametrization of complex systems

3.2.1. Project-team positioning

Clinical data are often sparse: we have few data per patient. The number of data is of the order of the number of parameters. In this context, a natural way to parametrize complex models with real world clinical data is to use a Bayesian approach, namely to try to find the distribution of the model parameters in the population, rather than to try to identify the parameters of every single patient. This approach has been pioneered in the 90's by the Nonmem software, and has been much improved thanks to Marc Lavielle in the 2000's. Refined statistical methods, called SAEM, have been tuned and implemented in commercial softwares like Monolix.

3.2.2. Recent results

The main problem when we try to parametrize clinical data using complex systems is the computational time. One single evaluation of the model can be costly, in particular if this model involves partial differential equations, and SAEM algorithm requires hundreds of thousands of single evaluations. The time cost is then too large, in particular because SAEM may not be parallelized.

To speed up the evaluation of the complex model, we replace it by an approximate one, or so called metamodel, constructed by interpolation of a small number of its values. We therefore combine the classical SAEM algorithm with an interpolation step, leading to a strong acceleration. Interpolation can be done through a precomputation step on a fixed grid, or through a more efficient kriging step. The interpolation grid or the kriging step may be improved during SAEM algorithm in an iterative way in order to get accurate evaluations of the complex system only in the domain of interest, namely near the clinical values [14],[15].

We applied these new algorithms to synthetic data and are currently using them on glioma data. We are also currently trying to prove the convergence of the corresponding algorithms. We will develop glioma applications in the next section.

Moreover E. Ollier in his phD developed new strategies to distinguish various populations within a SAEM algorithm [23].

We have two long standing collaborations with Sanofi and Servier on parametrization issues:

- Servier: during a four years contract, we modelled the pkpd of new drugs and also study the combination and optimization of chimiotherapies.
- Sanofi: during a eight years contract, Emmanuel Grenier wrote a complete software devoted to the study of the degradation of vaccine. This software is used worldwide by Sanofi R&D teams in order to investigate the degradation of existing or new vaccines and to study their behavior when they are heated. This software has been used on flu, dengue and various other diseases.

3.2.3. Collaborations

- Academic collaborations: A. Leclerc Samson (Grenoble University)
- Medical collaborations: Dr Ducray (Centre Léon Bérard, Lyon) and Dr Sujobert (Lyon Sud Hospital)
- Industrial contracts: we used parametrization and treatment improvement techniques for Servier (four years contract, on cancer drug modeling and optimization) and Sanofi (long standing collaboration)

3.3. Multiscale models in oncology

3.3.1. Project-team positioning

Cancer modeling is the major topic of several teams in France and Europe, including Mamba, Monc and Asclepios to quote only a few Inria teams. These teams try to model metastasis, tumoral growth, vascularisation through angiogenesis, or to improve medical images quality. Their approaches are based on dynamical systems, partial differential equations, or on special imagery techniques.

Numed focuses on the link between very simple partial differential equations models, like reaction diffusion models, and clinical data.

3.3.2. Results

During 2018 we developed new collaborations with the Centre Léon Bérard (Lyon), in particular on the following topics

- Barcoding of cells: thanks to recent techniques, it is possible to mark each cell with an individual barcode, and to follow its division and descendance. The analysis of such data requires probabilistic models, in particular to model experimental bias.
- Apoptosis: the question is to investigate whether the fate of neighboring cells influence the evolution of a given cell towards apoptosis, starting from videos of in vitro drug induced apoptosis.
- Dormance: Study of the dynamics of cells under immunotherapy, starting from experimental in vitro data.
- Colorectal cancer: In vitro study of the role of stem cells in drug resistance, in colorectal cancer.

3.3.3. Collaborations

- Centre Léon Bérard (in particular: Pr Puisieux, G. Ichim, M. Plateroni, S. Ortiz).

REO Project-Team

3. Research Program

3.1. Multiphysics modeling

In large vessels and in large bronchi, blood and air flows are generally supposed to be governed by the incompressible Navier-Stokes equations. Indeed in large arteries, blood can be supposed to be Newtonian, and at rest air can be modeled as an incompressible fluid. The cornerstone of the simulations is therefore a Navier-Stokes solver. But other physical features have also to be taken into account in simulations of biological flows, in particular fluid-structure interaction in large vessels and transport of sprays, particles or chemical species.

3.1.1. Fluid-structure interaction

Fluid-structure coupling occurs both in the respiratory and in the circulatory systems. We focus mainly on blood flows since our work is more advanced in this field. But the methods developed for blood flows could be also applied to the respiratory system.

Here "fluid-structure interaction" means a coupling between the 3D Navier-Stokes equations and a 3D (possibly thin) structure in large displacements.

The numerical simulations of the interaction between the artery wall and the blood flows raise many issues: (1) the displacement of the wall cannot be supposed to be infinitesimal, geometrical nonlinearities are therefore present in the structure and the fluid problem have to be solved on a moving domain (2) the densities of the artery walls and the blood being close, the coupling is strong and has to be tackled very carefully to avoid numerical instabilities, (3) "naive" boundary conditions on the artificial boundaries induce spurious reflection phenomena.

Simulation of valves, either at the outflow of the cardiac chambers or in veins, is another example of difficult fluid-structure problems arising in blood flows. In addition, very large displacements and changes of topology (contact problems) have to be handled in those cases.

Due to stability reasons, it seems impossible to successfully apply in hemodynamics the explicit coupling schemes used in other fluid-structure problems, like aeroelasticity. As a result, fluid-structure interaction in biological flows raise new challenging issues in scientific computing and numerical analysis : new schemes have to be developed and analyzed.

We have proposed and analyzed over the last few years several efficient fluid-structure interaction algorithms. This topic remains very active. We are now using these algorithms to address inverse problems in blood flows to make patient specific simulations (for example, estimation of artery wall stiffness from medical imaging).

3.1.2. Aerosol

Complex two-phase fluids can be modeled in many different ways. Eulerian models describe both phases by physical quantities such as the density, velocity or energy of each phase. In the mixed fluid-kinetic models, the biphasic fluid has one dispersed phase, which is constituted by a spray of droplets, with a possibly variable size, and a continuous classical fluid.

This type of model was first introduced by Williams [46] in the frame of combustion. It was later used to develop the Kiva code [36] at the Los Alamos National Laboratory, or the Hesione code [41], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [39], 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 [42]. The underlying idea consists in coupling different models (3D, 1D or 0D) with a decreasing level of accuracy, which is compensated by their decreasing level of computational complexity.

The research on this topic aims at providing a correct methodology and a mathematical and numerical framework for the simulation of blood flow in the whole cardiovascular system by means of a geometric multiscale approach. In particular, one of the main issues will be the definition of stable coupling strategies between 3D and reduced order models.

To model the arterial tree, a standard way consists of imposing a pressure or a flow rate at the inlet of the aorta, *i.e.* at the network entry. This strategy does not allow to describe important features as the overload in the heart caused by backward traveling waves. Indeed imposing a boundary condition at the beginning of the aorta artificially disturbs physiological pressure waves going from the arterial tree to the heart. The only way to catch this physiological behavior is to couple the arteries with a model of heart, or at least a model of left ventricle.

A constitutive law for the myocardium, controlled by an electrical command, has been developed in the CardioSense3D project [0]. One of our objectives is to couple artery models with this heart model.

A long term goal is to achieve 3D simulations of a system including heart and arteries. One of the difficulties of this very challenging task is to model the cardiac valves. To this purpose, we investigate a mix of arbitrary Lagrangian Eulerian and fictitious domain approaches or x-fem strategies, or simplified valve models based on an immersed surface strategy.

[0] http://www-sop.inria.fr/CardioSense3D/

3.2.2. Heart perfusion modeling

The heart is the organ that regulates, through its periodical contraction, the distribution of oxygenated blood in human vessels in order to nourish the different parts of the body. The heart needs its own supply of blood to work. The coronary arteries are the vessels that accomplish this task. The phenomenon by which blood reaches myocardial heart tissue starting from the blood vessels is called in medicine perfusion. The analysis of heart perfusion is an interesting and challenging problem. Our aim is to perform a three-dimensional dynamical numerical simulation of perfusion in the beating heart, in order to better understand the phenomena linked to perfusion. In particular the role of the ventricle contraction on the perfusion of the heart is investigated as well as the influence of blood on the solid mechanics of the ventricle. Heart perfusion in fact implies the interaction between heart muscle and blood vessels, in a sponge-like material that contracts at every heartbeat via the myocardium fibers.

Despite recent advances on the anatomical description and measurements of the coronary tree and on the corresponding physiological, physical and numerical modeling aspects, the complete modeling and simulation of blood flows inside the large and the many small vessels feeding the heart is still out of reach. Therefore, in order to model blood perfusion in the cardiac tissue, we must limit the description of the detailed flows at a given space scale, and simplify the modeling of the smaller scale flows by aggregating these phenomena into macroscopic quantities, by some kind of "homogenization" procedure. To that purpose, the modeling of the fluid-solid coupling within the framework of porous media appears appropriate.

Poromechanics is a simplified mixture theory where a complex fluid-structure interaction problem is replaced by a superposition of both components, each of them representing a fraction of the complete material at every point. It originally emerged in soils mechanics with the work of Terzaghi [45], and Biot [37] 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 [38]), albeit with *ad hoc* formulations for which compatibility with thermodynamics laws and incompressibility conditions is not established.

3.2.3. Tumor and vascularization

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

3.2.4. Respiratory tract modeling

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

SISTM Project-Team

3. Research Program

3.1. Mechanistic modelling

When studying the dynamics of a given marker, say the HIV concentration in the blood (HIV viral load), one can for instance use descriptive models summarising the dynamics over time in term of slopes of the trajectories [47]. These slopes can be compared between treatment groups or according to patients' characteristics. Another way for analysing these data is to define a mathematical model based on the biological knowledge of what drives HIV dynamics. In this case, it is mainly the availability of target cells (the CD4+ T lymphocytes), the production and death rates of infected cells and the clearance of the viral particles that impact the dynamics. Then, a mathematical model most often based on ordinary differential equations (ODE) can be written [40]. Estimating the parameters of this model to fit observed HIV viral load gave a crucial insight in HIV pathogenesis as it revealed the very short half-life of the virions and infected cells and therefore a very high turnover of the virus, making mutations a very frequent event [39].

Having a good mechanistic model in a biomedical context such as HIV infection opens doors to various applications beyond a good understanding of the data. Global and individual predictions can be excellent because of the external validity of a model based on main biological mechanisms. Control theory may serve for defining optimal interventions or optimal designs to evaluate new interventions [32]. Finally, these models can capture explicitly the complex relationship between several processes that change over time and may therefore challenge other proposed approaches such as marginal structural models to deal with causal associations in epidemiology [31].

Therefore, we postulate that this type of model could be very useful in the context of our research that is in complex biological systems. The definition of the model needs to identify the parameter values that fit the data. In clinical research this is challenging because data are sparse, and often unbalanced, coming from populations of subjects. A substantial inter-individual variability is always present and needs to be accounted as this is the main source of information. Although many approaches have been developed to estimate the parameters of non-linear mixed models [43], [50], [35], [41], [36], [49], the difficulty associated with the complexity of ODE models and the sparsity of the data leading to identifiability issues need further research.

3.2. High dimensional data

With the availability of omics data such as genomics (DNA), transcriptomics (RNA) or proteomics (proteins), but also other types of data, such as those arising from the combination of large observational databases (e.g. in pharmacoepidemiology or environmental epidemiology), high-dimensional data have became increasingly common. Use of molecular biological technics such as Polymerase Chain Reaction (PCR) allows for amplification of DNA or RNA sequences. Nowadays, microarray and Next Generation Sequencing (NGS) techniques give the possibility to explore very large portions of the genome. Furthermore, other assays have also evolved, and traditional measures such as cytometry or imaging have became new sources of big data. Therefore, in the context of HIV research, the dimension of the datasets has much grown in term of number of variables per individual than in term of number of included patients although this latter is also growing thanks to the multi-cohort collaborations such as CASCADE or COHERE organized in the EuroCoord network [0]. As an exemple, in a recent phase 1/2 clinical trial evaluating the safety and the immunological response to a dendritic cell-based HIV vaccine, 19 infected patients were included. Bringing together data on cell count, cytokine production, gene expression and viral genome change led to a 20 Go database [46]. This is far from big databases faced in other areas but constitutes a revolution in clinical research where clinical trials of hundred of patients sized few hundred of Ko at most. Therefore, more than the storage and calculation capacities, the challenge is the comprehensive analysis of these datasets.

[0] see online at http://www.eurocoord.net

The objective is either to select the relevant information or to summarize it for understanding or prediction purposes. When dealing with high dimensional data, the methodological challenge arises from the fact that datasets typically contain many variables, much more than observations. Hence, multiple testing is an obvious issue that needs to be taken into account [44]. Furthermore, conventional methods, such as linear models, are inefficient and most of the time even inapplicable. Specific methods have been developed, often derived from the machine learning field, such as regularization methods [48]. The integrative analysis of large datasets is challenging. For instance, one may want to look at the correlation between two large scale matrices composed by the transcriptome in the one hand and the proteome on the other hand [37]. The comprehensive analysis of these large datasets concerning several levels from molecular pathways to clinical response of a population of patients needs specific approaches and a very close collaboration with the providers of data that is the immunologists, the virologists, the clinicians...

XPOP Project-Team

3. Research Program

3.1. Scientific positioning

"Interfaces" is the defining characteristic of XPOP:

The interface between statistics, probability and numerical methods. Mathematical modelling of complex biological phenomena require to combine numerical, stochastic and statistical approaches. The CMAP is therefore the right place to be for positioning the team at the interface between several mathematical disciplines.

The interface between mathematics and the life sciences. The goal of XPOP is to bring the right answers to the right questions. These answers are mathematical tools (statistics, numerical methods, etc.), whereas the questions come from the life sciences (pharmacology, medicine, biology, etc.). This is why the point of XPOP is not to take part in mathematical projects only, but also pluridisciplinary ones.

The interface between mathematics and software development. The development of new methods is the main activity of XPOP. However, new methods are only useful if they end up being implemented in a software tool. On one hand, a strong partnership with Lixoft (the spin-off company who continue developing MONOLIX) allows us to maintaining this positioning. On the other hand, several members of the team are very active in the R community and develop widely used packages.

3.2. The mixed-effects models

Mixed-effects models are statistical models with both fixed effects and random effects. They are well-adapted to situations where repeated measurements are made on the same individual/statistical unit.

Consider first a single subject i of the population. Let $y_i = (y_{ij}, 1 \leq j \leq n_i)$ be the vector of observations for this subject. The model that describes the observations y_i is assumed to be a parametric probabilistic model: let $p_Y(y_i;\psi_i)$ be the probability distribution of y_i, where ψ_i is a vector of parameters.

In a population framework, the vector of parameters ψ_i is assumed to be drawn from a population distribution $p_\Psi(\psi_i;\theta)$ where θ is a vector of population parameters.

Then, the probabilistic model is the joint probability distribution

$$p(y_i,\psi_i;\theta) = p_Y(y_i|\psi_i)p_\Psi(\psi_i;\theta) \tag{19}$$

To define a model thus consists in defining precisely these two terms.

In most applications, the observed data y_i are continuous longitudinal data. We then assume the following representation for y_i:

$$y_{ij} = f(t_{ij},\psi_i) + g(t_{ij},\psi_i)\varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i. \tag{20}$$

Here, y_{ij} is the observation obtained from subject i at time t_{ij}. The residual errors (ε_{ij}) are assumed to be standardized random variables (mean zero and variance 1). The residual error model is represented by function g in model (2).

Function f is usually the solution to a system of ordinary differential equations (pharmacokinetic/pharmacodynamic models, etc.) or a system of partial differential equations (tumor growth, respiratory system, etc.). This component is a fundamental component of the model since it defines the prediction of the observed kinetics for a given set of parameters.

The vector of individual parameters ψ_i is usually function of a vector of population parameters $\psi_{\rm pop}$, a vector of random effects $\eta_i \sim \mathcal{N}(0, \Omega)$, a vector of individual covariates c_i (weight, age, gender, ...) and some fixed effects β.

The joint model of y and ψ depends then on a vector of parameters $\theta = (\psi_{\rm pop}, \beta, \Omega)$.

3.3. Computational Statistical Methods

Central to modern statistics is the use of probabilistic models. To relate these models to data requires the ability to calculate the probability of the observed data: the likelihood function, which is central to most statistical methods and provides a principled framework to handle uncertainty.

The emergence of computational statistics as a collection of powerful and general methodologies for carrying out likelihood-based inference made complex models with non-standard data accessible to likelihood, including hierarchical models, models with intricate latent structure, and missing data.

In particular, algorithms previously developed by POPIX for mixed effects models, and today implemented in several software tools (especially MONOLIX) are part of these methods:

- the adaptive Metropolis-Hastings algorithm allows one to sample from the conditional distribution of the individual parameters $p(\psi_i|y_i; c_i, \theta)$,
- the SAEM algorithm is used to maximize the observed likelihood $\mathcal{L}(\theta; y) = p(y; \theta)$,
- Importance Sampling Monte Carlo simulations provide an accurate estimation of the observed log-likelihood $\log(\mathcal{L}(\theta; y))$.

Computational statistics is an area which remains extremely active today. Recently, one can notice that the incentive for further improvements and innovation comes mainly from three broad directions: the high dimensional challenge, the quest for adaptive procedures that can eliminate the cumbersome process of tuning "by hand" the settings of the algorithms and the need for flexible theoretical support, arguably required by all recent developments as well as many of the traditional MCMC algorithms that are widely used in practice.

Working in these three directions is a clear objective for XPOP.

3.4. Markov Chain Monte Carlo algorithms

While these Monte Carlo algorithms have turned into standard tools over the past decade, they still face difficulties in handling less regular problems such as those involved in deriving inference for high-dimensional models. One of the main problems encountered when using MCMC in this challenging settings is that it is difficult to design a Markov chain that efficiently samples the state space of interest.

The Metropolis-adjusted Langevin algorithm (MALA) is a Markov chain Monte Carlo (MCMC) method for obtaining random samples from a probability distribution for which direct sampling is difficult. As the name suggests, MALA uses a combination of two mechanisms to generate the states of a random walk that has the target probability distribution as an invariant measure:

1. new states are proposed using Langevin dynamics, which use evaluations of the gradient of the target probability density function;
2. these proposals are accepted or rejected using the Metropolis-Hastings algorithm, which uses evaluations of the target probability density (but not its gradient).

Informally, the Langevin dynamics drives the random walk towards regions of high probability in the manner of a gradient flow, while the Metropolis-Hastings accept/reject mechanism improves the mixing and convergence properties of this random walk.

Several extensions of MALA have been proposed recently by several authors, including fMALA (fast MALA), AMALA (anisotropic MALA), MMALA (manifold MALA), position-dependent MALA (PMALA), ...

MALA and these extensions have demonstrated to represent very efficient alternative for sampling from high dimensional distributions. We therefore need to adapt these methods to general mixed effects models.

3.5. Parameter estimation

The Stochastic Approximation Expectation Maximization (SAEM) algorithm has shown to be extremely efficient for maximum likelihood estimation in incomplete data models, and particularly in mixed effects models for estimating the population parameters. However, there are several practical situations for which extensions of SAEM are still needed:

High dimensional model: a complex physiological model may have a large number of parameters (in the order of 100). Then several problems arise:

- when most of these parameters are associated with random effects, the MCMC algorithm should be able to sample, for each of the N individuals, parameters from a high dimensional distribution. Efficient MCMC methods for high dimensions are then required.
- Practical identifiability of the model is not ensured with a limited amount of data. In other words, we cannot expect to be able to properly estimate all the parameters of the model, including the fixed effects and the variance-covariance matrix of the random effects. Then, some random effects should be removed, assuming that some parameters do not vary in the population. It may also be necessary to fix the value of some parameters (using values from the literature for instance). The strategy to decide which parameters should be fixed and which random effects should be removed remains totally empirical. XPOP aims to develop a procedure that will help the modeller to take such decisions.

Large number of covariates: the covariate model aims to explain part of the inter-patient variability of some parameters. Classical methods for covariate model building are based on comparisons with respect to some criteria, usually derived from the likelihood (AIC, BIC), or some statistical test (Wald test, LRT, etc.). In other words, the modelling procedure requires two steps: first, all possible models are fitted using some estimation procedure (e.g. the SAEM algorithm) and the likelihood of each model is computed using a numerical integration procedure (e.g. Monte Carlo Importance Sampling); then, a model selection procedure chooses the "best" covariate model. Such a strategy is only possible with a reduced number of covariates, i.e., with a "small" number of models to fit and compare.

As an alternative, we are thinking about a Bayesian approach which consists of estimating simultaneously the covariate model and the parameters of the model in a single run. An (informative or uninformative) prior is defined for each model by defining a prior probability for each covariate to be included in the model. In other words, we extend the probabilistic model by introducing binary variables that indicate the presence or absence of each covariate in the model. Then, the model selection procedure consists of estimating and maximizing the conditional distribution of this sequence of binary variables. Furthermore, a probability can be associated to any of the possible covariate models.

This conditional distribution can be estimated using an MCMC procedure combined with the SAEM algorithm for estimating the population parameters of the model. In practice, such an approach can only deal with a limited number of covariates since the dimension of the probability space to explore increases exponentially with the number of covariates. Consequently, we would like to have methods able to find a small number of variables (from a large starting set) that influence certain parameters in populations of individuals. That means that, instead of estimating the conditional distribution of all the covariate models as described above, the algorithm should focus on the most likely ones.

Fixed parameters: it is quite frequent that some individual parameters of the model have no random component and are purely fixed effects. Then, the model may not belong to the exponential family anymore and the original version of SAEM cannot be used as it is. Several extensions exist:

- introduce random effects with decreasing variances for these parameters,
- introduce a prior distribution for these fixed effects,
- apply the stochastic approximation directly on the sequence of estimated parameters, instead of the sufficient statistics of the model.

None of these methods always work correctly. Furthermore, what are the pros and cons of these methods is not clear at all. Then, developing a robust methodology for such model is necessary.

Convergence toward the global maximum of the likelihood: convergence of SAEM can strongly depend on thie initial guess when the observed likelihood has several local maxima. A kind of simulated annealing version of SAEM was previously developed and implemented in MONOLIX. The method works quite well in most situations but there is no theoretical justification and choosing the settings of this algorithm (i.e. how the temperature decreases during the iterations) remains empirical. A precise analysis of the algorithm could be very useful to better understand why it "works" in practice and how to optimize it.

Convergence diagnostic: Convergence of SAEM was theoretically demonstrated under very general hypothesis. Such result is important but of little interest in practice at the time to use SAEM in a finite amount of time, i.e. in a finite number of iterations. Some qualitative and quantitative criteria should be defined in order to both optimize the settings of the algorithm, detect a poor convergence of SAEM and evaluate the quality of the results in order to avoid using them unwisely.

3.6. Model building

Defining an optimal strategy for model building is far from easy because a model is the assembled product of numerous components that need to been evaluated and perhaps improved: the structural model, residual error model, covariate model, covariance model, etc.

How to proceed so as to obtain the best possible combination of these components? There is no magic recipe but an effort will be made to provide some qualitative and quantitative criteria in order to help the modeller for building his model.

The strategy to take will mainly depend on the time we can dedicate to building the model and the time required for running it. For relatively simple models for which parameter estimation is fast, it is possible to fit many models and compare them. This can also be done if we have powerful computing facilities available (e.g., a cluster) allowing large numbers of simultaneous runs.

However, if we are working on a standard laptop or desktop computer, model building is a sequential process in which a new model is tested at each step. If the model is complex and requires significant computation time (e.g., when involving systems of ODEs), we are constrained to limit the number of models we can test in a reasonable time period. In this context, it also becomes important to carefully choose the tasks to run at each step.

3.7. Model evaluation

Diagnostic tools are recognized as an essential method for model assessment in the process of model building. Indeed, the modeler needs to confront "his" model with the experimental data before concluding that this model is able to reproduce the data and before using it for any purpose, such as prediction or simulation for instance.

The objective of a diagnostic tool is twofold: first we want to check if the assumptions made on the model are valid or not ; then, if some assumptions are rejected, we want to get some guidance on how to improve the model.

As is the usual case in statistics, it is not because this "final" model has not been rejected that it is necessarily the "true" one. All that we can say is that the experimental data does not allow us to reject it. It is merely one of perhaps many models that cannot be rejected.

Model diagnostic tools are for the most part graphical, i.e., visual; we "see" when something is not right between a chosen model and the data it is hypothesized to describe. These diagnostic plots are usually based on the empirical Bayes estimates (EBEs) of the individual parameters and EBEs of the random effects: scatterplots of individual parameters versus covariates to detect some possible relationship, scatterplots of pairs of random effects to detect some possible correlation between random effects, plot of the empirical distribution of the random effects (boxplot, histogram,...) to check if they are normally distributed, ...

The use of EBEs for diagnostic plots and statistical tests is efficient with rich data, i.e. when a significant amount of information is available in the data for recovering accurately all the individual parameters. On the contrary, tests and plots can be misleading when the estimates of the individual parameters are greatly shrunk.

We propose to develop new approaches for diagnosing mixed effects models in a general context and derive formal and unbiased statistical tests for testing separately each feature of the model.

3.8. Missing data

The ability to easily collect and gather a large amount of data from different sources can be seen as an opportunity to better understand many processes. It has already led to breakthroughs in several application areas. However, due to the wide heterogeneity of measurements and objectives, these large databases often exhibit an extraordinary high number of missing values. Hence, in addition to scientific questions, such data also present some important methodological and technical challenges for data analyst.

Missing values occur for a variety of reasons: machines that fail, survey participants who do not answer certain questions, destroyed or lost data, dead animals, damaged plants, etc. Missing values are problematic since most statistical methods can not be applied directly on a incomplete data. Many progress have been made to properly handle missing values. However, there are still many challenges that need to be addressed in the future, that are crucial for the users.

- State of arts methods often consider the case of continuous or categorical data whereas real data are very often mixed. The idea is to develop a multiple imputation method based on a specific principal component analysis (PCA) for mixed data. Indeed, PCA has been used with success to predict (impute) the missing values. A very appealing property is the ability of the method to handle very large matrices with large amount of missing entries.
- The asymptotic regime underlying modern data is not any more to consider that the sample size increases but that both number of observations and number of variables are very large. In practice first experiments showed that the coverage properties of confidence areas based on the classical methods to estimate variance with missing values varied widely. The asymptotic method and the bootstrap do well in low-noise setting, but can fail when the noise level gets high or when the number of variables is much greater than the number of rows. On the other hand, the jackknife has good coverage properties for large noisy examples but requires a minimum number of variables to be stable enough.
- Inference with missing values is usually performed under the assumption of "Missing at Random" (MAR) values which means that the probability that a value is missing may depend on the observed data but does not depend on the missing value itself. In real data and in particular in data coming from clinical studies, both "Missing Non at Random" (MNAR) and MAR values occur. Taking into account in a proper way both types of missing values is extremely challenging but is worth investigating since the applications are extremely broad.

It is important to stress that missing data models are part of the general incomplete data models addressed by XPOP. Indeed, models with latent variables (i.e. non observed variables such as random effects in a mixed effects model), models with censored data (e.g. data below some limit of quantification) or models with dropout mechanism (e.g. when a subject in a clinical trial fails to continue in the study) can be seen as missing data models.