Activity Report 2016

Project-Team ABS

Algorithms, Biology, Structure

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Computational Biology

Table of contents

Project-Team ABS

Creation of the Project-Team: 2008 July 01

Keywords:

Computer Science and Digital Science:

2.5. - Software engineering
3.3.2. - Data mining
3.4.1. - Supervised learning
3.4.2. - Unsupervised learning
6.1.4. - Multiscale modeling
6.2.4. - Statistical methods
6.2.8. - Computational geometry and meshes
7.2. - Discrete mathematics, combinatorics
7.5. - Geometry, Topology
7.9. - Graph theory
8.2. - Machine learning

Other Research Topics and Application Domains:

1.1.1. - Structural biology
1.1.7. - Immunology
1.1.9. - Bioinformatics

1. Members

Research Scientists

Frédéric Cazals [Team leader, Inria, Senior Researcher, HDR]
Dorian Mazauric [Inria, Researcher]

PhD Students

Augustin Chevallier [Univ. Nice, PhD Student]
Simon Marillet [INRA, PhD Student, until December 2015]
Romain Tetley [Univ. Nice, PhD Student]
Tom Dreyfus [Redant labs - Inria, contractor]
Neva Durand [Baylor college of medicine, Visiting scientist]

Post-Doctoral Fellow

Rémi Watrigant [Inria, Post-Doctoral Fellow, from Nov 2016]

Administrative Assistant

Florence Barbara [Inria, Assistant]

2. Overall Objectives

2.1. Overall Objectives

Computational Biology and Computational Structural Biology. Understanding the lineage between species and the genetic drift of genes and genomes, apprehending the control and feed-back loops governing the behavior of a cell, a tissue, an organ or a body, and inferring the relationship between the structure of biological (macro)-molecules and their functions are amongst the major challenges of modern biology. The investigation of these challenges is supported by three types of data: genomic data, transcription and expression data, and structural data.

Genetic data feature sequences of nucleotides on DNA and RNA molecules, and are symbolic data whose processing falls in the realm of Theoretical Computer Science: dynamic programming, algorithms on texts and strings, graph theory dedicated to phylogenetic problems. Transcription and expression data feature evolving concentrations of molecules (RNAs, proteins, metabolites) over time, and fit in the formalism of discrete and continuous dynamical systems, and of graph theory. The exploration and the modeling of these data are covered by a rapidly expanding research field termed *systems biology*. Structural data encode informations about the 3D structures of molecules (nucleic acids (DNA, RNA), proteins, small molecules) and their interactions, and come from three main sources: X ray crystallography, NMR spectroscopy, cryo Electron Microscopy. Ultimately, structural data should expand our understanding of how the structure accounts for the function of macro-molecules – one of the central questions in structural biology. This goal actually subsumes two equally difficult challenges, which are *folding* – the process through which a protein adopts its 3D structure, and *docking* – the process through which two or several molecules assemble. Folding and docking are driven by non covalent interactions, and for complex systems, are actually inter-twined [48]. Apart from the bio-physical interests raised by these processes, two different application domains are concerned: in fundamental biology, one is primarily interested in understanding the machinery of the cell; in medicine, applications to drug design are developed.

Modeling in Computational Structural Biology. Acquiring structural data is not always possible: NMR is restricted to relatively small molecules; membrane proteins do not crystallize, etc. As a matter of fact, the order of magnitude of the number of genomes sequenced is of the order of one thousand, which results in circa one million of genes recorded in the manually curated Swiss-Prot database. On the other hand, the Protein Data Bank contains circa 90,000 structures. Thus, the paucity of structures with respect to the known number of genes calls for modeling in structural biology, so as to foster our understanding of the structure-to-function relationship.

Ideally, bio-physical models of macro-molecules should resort to quantum mechanics. While this is possible for small systems, say up to 50 atoms, large systems are investigated within the framework of the Born-Oppenheimer approximation which stipulates the nuclei and the electron cloud can be decoupled. Example force fields developed in this realm are AMBER, CHARMM, OPLS. Of particular importance are Van der Waals models, where each atom is modeled by a sphere whose radius depends on the atom chemical type. From an historical perspective, Richards [46], [35] and later Connolly [31], while defining molecular surfaces and developing algorithms to compute them, established the connexions between molecular modeling and geometric constructions. Remarkably, a number of difficult problems (e.g. additively weighted Voronoi diagrams) were touched upon in these early days.

The models developed in this vein are instrumental in investigating the interactions of molecules for which no structural data is available. But such models often fall short from providing complete answers, which we illustrate with the folding problem. On one hand, as the conformations of side-chains belong to discrete sets (the so-called rotamers or rotational isomers) [37], the number of distinct conformations of a poly-peptidic chain is exponential in the number of amino-acids. On the other hand, Nature folds proteins within time scales ranging from milliseconds to hours, while time-steps used in molecular dynamics simulations are of the order of the femto-second, so that biologically relevant time-scales are out reach for simulations. The fact that Nature avoids the exponential trap is known as Levinthal's paradox. The intrinsic difficulty of problems calls for models exploiting several classes of informations. For small systems, *ab initio* models can be built from first principles. But for more complex systems, *homology* or template-based models integrating a variable amount of knowledge acquired on similar systems are resorted to.

The variety of approaches developed are illustrated by the two community wide experiments CASP (*Critical Assessment of Techniques for Protein Structure Prediction*; http://predictioncenter.org) and CAPRI (*Critical Assessment of Prediction of Interactions*; http://capri.ebi.ac.uk), which allow models and prediction algorithms to be compared to experimentally resolved structures.

As illustrated by the previous discussion, modeling macro-molecules touches upon biology, physics and chemistry, as well as mathematics and computer science. In the following, we present the topics investigated within ABS.

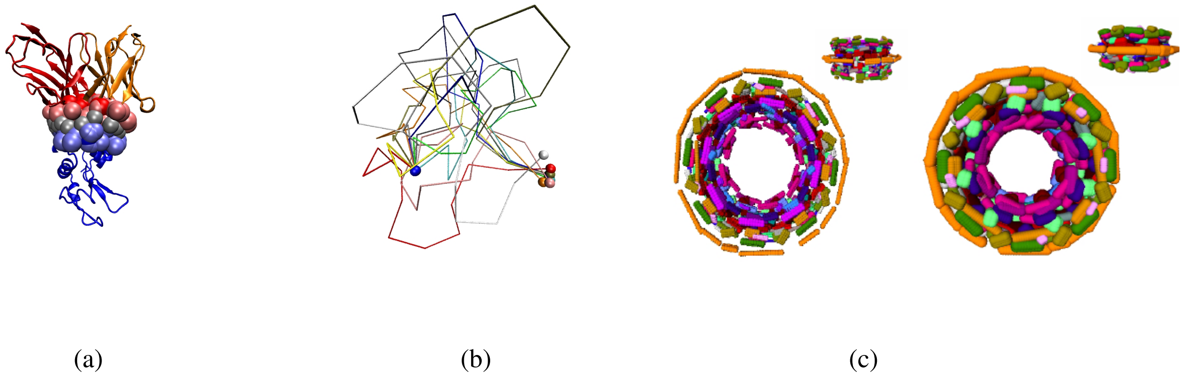

Figure 1. ***Geometric constructions in computational structural biology.*** *(a) An antibody-antigen complex, with interface atoms identified by our Voronoi based interface model. This model is instrumental in mining correlations between structural and biological as well as biophysical properties of protein complexes [12]. (b) A diverse set of conformations of a backbone loop, selected thanks to a geometric optimization algorithm [6]. Such conformations are used by mean field theory based docking algorithms. (c) A toleranced model (TOM) of the nuclear pore complex, visualized at two different scales [9]. The parameterized family of shapes coded by a TOM is instrumental to identify stable properties of the underlying macro-molecular system.*

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 [1], the remaining ones being protein - protein or protein - drug complexes. These structures feature what Nature does – up to the bias imposed by the experimental conditions inherent to structure elucidation, and are of special interest to investigate non-covalent contacts in biological complexes. More precisely, given two proteins defining a complex, interface atoms are defined as the atoms of one protein *interacting* with atoms of the second one. Understanding the structure of interfaces is central to understand biological complexes and thus the function of biological molecules [48]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [51]. Current investigations follow two routes. From the experimental perspective [34], 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

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

[45]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [40].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change [2], or a function stating that a complex is more stable than another one, in which case the value returned is a score and not an energy. Borrowing to statistical mechanics [29], 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 [49], [36]. 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 [41]. 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.

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 [30]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [50], 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 [33].

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 [44]. 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 [28]. 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 [27], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

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

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 [26], 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 [26]. 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 [3]. Understanding correlations is of special interest to predict the folding pathway that leads a protein towards its native state. A similar discussion holds for the case of partners within a complex, for example in the third step of the *diffusion - conformer selection - induced fit* complex formation model.

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

At the side-chain level, the question of improving rotamer libraries is still of interest [32]. 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) [47]. 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 [43], to Morse theory [38] and to analysis of meta-stable states of time series [39] have been proposed.

3.5. Algorithmic foundations

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

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

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.

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 [42].

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 [7]. 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.

4. Highlights of the Year

4.1. Highlights of the Year

In 2016, several achievements are worth noticing in three realms, namely in computer science, computational structural biology, and software.

4.1.1. Computer Science

► **Optimal transportation problems with connectivity constraints**
Reference: [21]

In a nutshell: Optimal transportation theory provides a rich framework to compare *measures*, both in the continuous and discrete settings. In this work, we study generalization of discrete transportation problems, when the supply and demand nodes are endowed with a graph structure; due to these constraints, our study focuses on transport plans respecting selected connectivity constrains. Our contributions encompass a formalization of these problems, as well as hardness results and heuristic algorithms.

Assessment: To the best of our knowledge, this work is the first one focusing on transport plans with connectivity constraints. One of the key applications targeted is the comparison of potential energy landscapes (PEL) in biophysics. Our algorithms provide a novel way to compare PEL, a topic overlooked so far.

► **Clustering stability revealed by matchings between clusters of clusters**
Reference: [22]

In a nutshell: Clustering is a fundamental problem in data science, yet, the variety of clustering methods and their sensitivity to parameters make clustering hard. This work provides a new tier of methods to compare two clusterings, by computing meta-clusters within each clustering– a meta-cluster is a group of clusters, together with a matching between these.

Assessment: Our methods will help assess the coherence between two clusterings, in two respects: by stressing the (lack of) stability of clustering while varying the parameters of a given algorithm, and by allowing a detailed comparisons of various algorithms.

4.1.2. Computational Structural Biology

► **Novel structural parameters of Ig-Ag complexes yield a quantitative description of interaction specificity and binding affinity**
Reference: [23]

In a nutshell: Understanding the specificity of antibodies for the targeted antigens, and predicting the affinity an antibody - antigen complexes is a central question in structural immunology. Using novel parameters acting as proxys for important biophysical quantities, we obtained affinity predictions of unprecedented accuracy, and were able to provide a quantitative explanation for the specific role of so-called *complementarity determining regions* – in particular CDR3 of heavy chains. See details in section 6.1.2.

Assessment: Our affinity predictions are the most accurate known to date, and show that for certain classes of IG - Ag complexes, the affinity prediction problem may be solved from databases of high resolution crystal structures.

► **Energy landscapes and persistent minima**
Reference: [15]

In a nutshell: Potential energy landscapes (PEL) of molecular systems are complex high-dimensional height functions. In this work, we introduced several tools from graph theory, optimization, and computational topology, so as to identify prominent features of PEL – prosaically distinguishing the signal from the noise. See details in section 6.3.1.

Assessment: Our work calls for important developments in two directions. The first one is concerned with the *calibration / learning* of features of PEL. The second one is the systematic comparison of force fields used in biophysics, as from current knowledge, deciding which force field is best for a given task or system is an open issue.

▶ **Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes**
Reference: [18]

In a nutshell: We developed a novel exploration algorithm for high-dimensional non convex (potential) energy functions used in biophysics. Our algorithm exploits the ability of *basin hopping* to locate low-lying local minima, and that of *rapidly exploring random tree* to foster the exploration of yet unexplored regions. See details in section 6.3.2.

Assessment: Our exploration algorithm outperform the two classical algorithms it is derived from. To strike a major impact, though, our exploration strategy needs to be complemented by enhanced thermodynamic sampling algorithms, able to bridge the gap between structures on the one hand, and thermodynamics / dynamics on the other hand.

4.1.3. Software

▶ **The Structural Bioinformatics Library**
Reference: [20]

In a nutshell: The SBL was released in 2015. In 2016, two important milestones were achieved, with the addition of several important packages, notably geared towards the generation and the analysis of conformational ensembles, and the publication of [20]–to appear in Bioinformatics.

Assessment: As outlined by the reviewers of [20], the SBL is to the best of our knowledge the first library proposing a coherent framework, in terms of algorithms, data structures and biophysical models, to tackle the most important problems in structural bioinformatics. Our paper presenting the SBL being in press as of December 2016, statistics on users and downloads will be reported in the 2017 activity report.

5. New Software and Platforms

5.1. The Structural Bioinformatics Library

5.1.1. The SBL : overview

The `SBL` (http://sbl.inria.fr) is a generic C++/python library providing algorithms and applications to solve complex problems in computational structural biology (CSB). [20].

For Biologists, the key advantages are:

- comprehensive in silico environment providing software applications,
- answering complex bio-physical problems (modeling interfaces and contacts, modeling the flexibility of proteins, and modeling macro-molecular assemblies),
- in a robust, fast and reproducible way.

For Developers, the striking facts are:

- broad C++/python toolbox,
- with modular design and careful specifications,
- fostering the development of complex applications.

5.1.2. The SBL : rationale and design

Software development generally faces a dichotomy, with on the one hand generic libraries providing methods of ubiquitous interest, and on the other hand application driven libraries targeting specific application areas. Libraries in the former category typically provide state-of-the art low level algorithms carefully specified, at the detriment of high level applications. Libraries in the latter category are generally high level and user-friendly, but the lack of formalism often makes it difficult to couple them to low level algorithms with formal specifications. The `SBL` ambitions to reconcile both software development philosophies, based on an advanced design suited for all classes of users and developers.

In terms of high-level operations, the `SBL` provides various applications revolving around the problem of understanding the relationship between the structure and the function of macro-molecules and their complexes (see below). In terms of low-level operations, the design of the `SBL` is meant to accommodate both the variety of models coding the physical and chemical properties of macro-molecular systems (models based on unions of balls such as van der Walls models or solvent accessible models, or models based on conformations and conformational ensembles), as well as the variety of operations (geometric, topological, and combinatorial) undertaken on these models.

More precisely, the `SBL` consists of the following software components, detailed below:

- `SBL-APPLICATIONS`**:** high level applications solving specific applied problems.
- `SBL-CORE`**:** low-level generic C++ classes templated by traits classes specifying C++ concepts [4].
- `SBL-MODELS`**:** C++ *models* matching the C++ concepts required to instantiate classes from `SBL-CORE`.
- `SBL-MODULES`**:** C++ classes instantiating classes from the `SBL-CORE` with specific biophysical models from `SBL-MODELS`. A module may be seen as a black box transforming an input into an output. With modules, an application workflow consists of interconnected modules.

5.1.3. The SBL for end-users: SBL-APPLICATIONS

End users will find in the `SBL` portable applications running on all platforms (Linux, MacOS, Windows). These applications split into the following categories:

- **Space Filling Models:** applications dealing with molecular models defined by unions of balls.
- **Conformational Analysis:** applications dealing with molecular flexibility.
- **Large assemblies:** applications dealing with macro-molecular assemblies involving from tens to hundreds of macro-molecules.
- **Data Analysis:** applications providing novel data analysis - statistical analysis tools.
- **Data Management:** applications to handle input data and results, using standard tools revolving around the XML file format (in particular the XPath query language). These tools allow automating data storage, parsing and retrieval, so that upon running calculations with applications, statistical analysis and plots are a handful of python lines away.

5.1.4. The SBL for developers: SBL-CORE, SBL-MODELS and SBL-MODULES

The `SBL` makes it easy to develop novel high-level applications, by providing high level ready to use C++ classes instantiating various biophysical models.

In particular, modules allow the development of applications without the burden of instantiating low level classes. In fact, once modules are available, designing an application merely consists of connecting modules.

[4] The design has been guided by that used in the Computational Geometry Algorithm Library (CGAL), see http://www.cgal.org

5.1.5. The SBL for low-level developers and contributors: SBL-CORE, and SBL-MODELS

Low level developments may use classes from / contribute classes to SBL-CORE and SBL-MODELS. In fact, such developments are equivalent to those based upon C++ libraries such as CGAL (http://www.cgal.org/) or boost C++ libraries (http://www.boost.org/). It should be noticed that the SBL heavily relies on these libraries. The SBL-CORE is organized into four sub-sections:

- CADS : Combinatorial Algorithms and Data Structures.
- GT : Computational geometry and computational topology.
- CSB : Computational Structural Biology.
- IO : Input / Output.

It should also be stressed that these packages implement algorithms not available elsewhere, or available in a non-generic guise. Due to the modular structure of the library, should valuable implementations be made available outside the SBL (e.g. in CGAL or boost), a substitution may occur.

5.1.6. Interoperability

The SBL is interoperable with existing molecular modeling systems, at several levels:

- At the library level, our state-of-the-art algorithms (e.g. the computation of molecular surfaces and volumes) can be integrated within existing software by instantiating the required classes from SBL-CORE, or using the adequate modules.
- At the application level, our applications can easily be integrated within processing pipelines, since the format used for input and output are standard ones. (For input, the PDB format can always be used. For output, our applications generate XML files.)
- Finally, for visualization purposes, our applications generate outputs for the two reference molecular modeling environments, namely Visual Molecular Dynamics (http://www.ks.uiuc.edu/Research/vmd/) and Pymol (http://www.pymol.org/).

5.1.7. Releases, distribution, and license

The SBL is released under a proprietary open source license, see http://sbl.inria.fr/license/.

The source code is distributed from http://sbl.inria.fr, using tarballs and a git repository. Bugzilla is used to handle user's feedback and bug tracking.

6. New Results

6.1. Modeling interfaces and contacts

Keywords: docking, scoring, interfaces, protein complexes, Voronoi diagrams, arrangements of balls.

6.1.1. Predicting binding poses and affinities for protein - ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation

Participants: Frédéric Cazals, Simon Marillet.

In collaboration with Sergei Grudinin, Maria Kadukova and Andreas Eisenbarth (Univ. Grenoble Alpes / CNRS / Inria, France).

The 2015 D3R Grand Challenge provided an opportunity to test our new model for the binding free energy of small molecules [17], as well as to assess our protocol to predict binding poses for protein-ligand complexes. Our pose predictions were ranked 3-9 for the HSP90 dataset, depending on the assessment metric. For the MAP4K dataset the ranks are very dispersed and equal to 2-35, depending on the assessment metric, which does not provide any insight into the accuracy of the method. The main success of our pose prediction protocol was the re-scoring stage using the recently developed Convex-PL potential. We make a thorough analysis of our docking predictions and discuss the effect of the choice of rigid receptor templates, the number of flexible residues in the binding pocket, the binding pocket size, and the subsequent re-scoring.

However, the main challenge was to predict experimentally determined binding affinities for two blind test sets. Our affinity prediction model consisted of two terms, a pairwise-additive enthalpy, and a non pairwise-additive entropy. We trained the free parameters of the model with a regularized regression using affinity and structural data from the PDBBind database. Our model performed very well on the training set, however, failed on the two test sets. We explain the drawback and pitfalls of our model, in particular in terms of relative coverage of the test set by the training set and missed dynamical properties from crystal structures, and discuss different routes to improve it.

6.1.2. *Novel structural parameters of Ig-Ag complexes yield a quantitative description of interaction specificity and binding affinity*

Participants: Frédéric Cazals, Simon Marillet.

In collaboration with Pierre Boudinot (INRA Jouy-en-Josas) and M-P. Lefranc (University of Montpellier 2).

Antibody-antigen complexes challenge our understanding, as analyses to date failed to unveil the key determinants of binding affinity and interaction specificity. In this work [23], we partially fill this gap based on novel quantitative analyses using two standardized databases, the IMGT/3Dstructure-DB and the structure affinity benchmark.

First, we introduce a statistical analysis of interfaces which enables the classification of ligand types (protein, peptide, chemical; cross-validated classification error of 9.6%), and yield binding affinity predictions of unprecedented accuracy (median absolute error of 0.878 kcal/mol). Second, we exploit the contributions made by CDRs in terms of position at the interface and atomic packing properties to show that in general, VH CDR3 and VL CDR3 make dominant contributions to the binding affinity, a fact also shown to be consistent with the enthalpy - entropy compensation associated with pre-configuration of CDR3. Our work suggests that the affinity prediction problem could be solved from databases of high resolution crystal structures of complexes with known affinity.

6.2. Modeling macro-molecular assemblies

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

No new result on this topic in 2016.

6.3. Modeling the flexibility of macro-molecules

Keywords: protein, flexibility, collective coordinate, conformational sampling dimensionality reduction.

6.3.1. *Energy landscapes and persistent minima*

Participants: Frédéric Cazals, Dorian Mazauric.

In collaboration with David Wales and Joanne Carr, from Cambridge University (UK).

In this work [15], we consider a coarse-graining of high-dimensional potential energy landscapes based upon persistences—which correspond to lowest barrier heights to lower-energy minima. Persistences can be calculated efficiently for local minima in kinetic transition networks that are based on stationary points of the prevailing energy landscape. The networks studied here represent peptides, proteins, nucleic acids, an atomic cluster, and a glassy system. Minima with high persistence values are likely to represent some form of alternative structural morphology, which, if appreciably populated at the prevailing temperature, could compete with the global minimum (defined as in- finitely persistent). Threshold values on persistences (and in some cases equilibrium occupation probabilities) have therefore been used in this work to select subsets of minima, which were then analysed to see how well they can represent features of the full network. Simplified disconnectivity graphs showing only the selected minima can convey the funnelling (including any multiple-funnel) characteristics of the corresponding full graphs. The effect of the choice of persistence threshold on the reduced disconnectivity graphs was considered for a system with a hierarchical, glassy landscape. Sets of persistent minima were also found to be useful in comparing networks for the same system sampled under different conditions, using minimum oriented spanning forests.

6.3.2. Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes

Participants: Frédéric Cazals, Tom Dreyfus, Andrea Roth.

In collaboration with Charles Robert (IBPC / CNRS, Paris, France).

The number of local minima of the potential energy landscape (PEL) of molecular systems generally grows exponentially with the number of degrees of freedom, so that a crucial property of PEL exploration algorithms is their ability to identify local minima which are low lying and diverse. In this work [18], we present a new exploration algorithm, retaining the ability of basin hopping (BH) to identify local minima, and that of transition based rapidly exploring random trees (T-RRT) to foster the exploration of yet unexplored regions. This ability is obtained by interleaving calls to the extension procedures of BH and T-RRT, and we show tuning the balance between these two types of calls allows the algorithm to focus on low lying regions. Computational efficiency is obtained using state-of-the art data structures, in particular for searching approximate nearest neighbors in metric spaces. We present results for the BLN69, a protein model whose conformational space has dimension 207 and whose PEL has been studied exhaustively. On this system, we show that the propensity of our algorithm to explore low lying regions of the landscape significantly outperforms those of BH and T-RRT.

6.4. Algorithmic foundations

Keywords: computational geometry, Computational topology, Voronoi diagrams, α-shapes, Morse theory, graph algorithm, combinatorial optimization, statistical learning.

6.4.1. The Structural Bioinformatics Library: modeling in biomolecular science and beyond

Participants: Frédéric Cazals, Tom Dreyfus.

Software in structural bioinformatics has mainly been application driven. To favor practitioners seeking off-the-shelf applications, but also developers seeking advanced building blocks to develop novel applications, we undertook the design of the Structural Bioinformatics Library (SBL, http://sbl. inria.fr) [20], a generic C++/python cross-platform software library targeting complex problems in structural bioinformatics. Its tenet is based on a modular design offering a rich and versatile framework allowing the development of novel applications requiring well specified complex operations, without compromising robustness and performances.

The SBL involves four software components (1-4 thereafter). For end-users, the SBL provides ready to use, state-of-the-art (1) applications to handle molecular models defined by unions of balls, to deal with molecular flexibility, to model macro-molecular assemblies. These tools can also be combined to tackle integrated analysis problems. For developers, the SBL provides a broad C++ toolbox with modular design, involving (2) core algorithms, (3) biophysical models, and (4) modules, the latter being especially suited to develop novel applications. The SBL comes with a thorough documentation consisting of user and reference manuals, and a bugzilla platform to handle community feedback.

The SBL is available from http://sbl.inria.fr.

6.4.2. Optimal transportation problems with connectivity constraints

Participants: Frédéric Cazals, Dorian Mazauric.

The earth mover distance (EMD) or the Mallows distance are example optimal transportation (OT) problems reducing to linear programs. In this work [21], we study a generalization of these problems when the supply and demand nodes are the vertices of two graphs called the supply and the demand graphs. The novel problems embed connectivity constraints in the transport plans computed, using a Lipschitz-like condition involving distances between certain subgraphs of the supply graph and certain subgraphs of the demand graph. More precisely, we make three contributions.

First, we formally introduce two optimal transportation problems generalizing EMD, namely Minimum-cost under flow, transport size, and connectivity constraints problem (problem EMD-FCC) and Maximum-flow under cost, transport size, and connectivity constraints problem (problem EMD-CCC). We prove that problems EMD-CCC and EMD-FCC are NP-complete, and that EMD-FCC is hard to approximate within any given constant. Second, we develop a greedy heuristic algorithm returning admissible solutions, of time complexity $O(n^3m^2)$ with n and m the numbers of vertices of the supply and demand graphs, respectively. Third, on the experimental side, we apply our novel OT algorithms for two applications, namely the comparison of clusterings, and the analysis of so-called potential energy landscapes in molecular science. These experiments show that optimizing the transport plan and respecting connectivity constraint can be competing objectives. Implementations of our algorithms are available in the Structural Bioinformatics Library at http://sbl.inria.fr.

6.4.3. Clustering stability revealed by matchings between clusters of clusters

Participants: Frédéric Cazals, Dorian Mazauric, Romain Tetley.

Clustering is a fundamental problem in data science, yet, the variety of clustering methods and their sensitivity to parameters make clustering hard. To analyze the stability of a given clustering algorithm while varying its parameters, and to compare clusters yielded by different algorithms, several comparison schemes based on matchings, information theory and various indices (Rand, Jaccard) have been developed. In this work [22], we go beyond these by providing a novel class of methods computing meta-clusters within each clustering– a meta-cluster is a group of clusters, together with a matching between these. Altogether, these pieces of information help assessing the coherence between two clusterings.

More specifically, let the intersection graph of two clusterings be the edge-weighted bipartite graph in which the nodes represent the clusters, the edges represent the non empty intersection between two clusters, and the weight of an edge is the number of common items. We introduce the so-called (k,D) and D-family-matching problems on intersection graphs, with k the number of meta-clusters and D the upper-bound on the diameter of the graph induced by the clusters of any meta-cluster. First we prove hardness and inapproximability results. Second, we design exact polynomial time dynamic programming algorithms for some classes of graphs (in particular trees). Then, we prove efficient (exact, approximation, and heuristic) algorithms, based on spanning trees, for general graphs. Practically, we present extensive experiments in two directions. First, we illustrate the ability of our algorithms to identify relevant meta-clusters between a given clustering and an edited version of it. Second, we show how our methods can be used to identify notorious instabilities of the k-means algorithm.

6.4.4. Experimental evaluation of a branch and bound algorithm for computing pathwidth

Participant: Dorian Mazauric.

In collaboration with David Coudert and Nicolas Nisse (COATI project-team, Université Côte D'Azur, Inria, I3S / CNRS).

Path-decompositions of graphs are an important ingredient of dynamic programming algorithms for solving efficiently many NP-hard problems. Therefore, computing the pathwidth and associated path-decomposition of graphs has both a theoretical and practical interest. In [16], we design a Branch and Bound algorithm that computes the exact pathwidth of graphs and a corresponding path-decomposition. Our main contribution consists of several non-trivial techniques to reduce the size of the input graph (pre-processing) and to cut the exploration space during the search phase of the algorithm. We evaluate experimentally our algorithm by comparing it to existing algorithms of the literature. It appears from the simulations that our algorithm offers a significant gain with respect to previous work. In particular, it is able to compute the exact pathwidth of any graph with less than 60 nodes in a reasonable running-time (≤ 10 minutes on a standard laptop). Moreover, our algorithm achieves good performance when used as a heuristic (i.e., when returning best result found within bounded time-limit). Our algorithm is not restricted to undirected graphs since it actually computes the directed pathwidth which generalizes the notion of pathwidth to digraphs.

6.4.5. Extracting the core structural connectivity network: guaranteeing network connectedness through a graph-theoretical approach

Participant: Dorian Mazauric.

In collaboration with Demian Wassermann, Guillermo Gallardo-Diez and Rachid Deriche (ATHENA project-team, Université Côte d'Azur, Inria).

In this work [19], we present a graph-theoretical algorithm to extract the connected core structural connectivity network of a subject population. Extracting this core common network across subjects is a main problem in current neuroscience. Such network facilitates cognitive and clinical analyses by reducing the number of connections that need to be explored. Furthermore, insights into the human brain structure can be gained by comparing core networks of different populations. We show that our novel algorithm has theoretical and practical advantages. First, contrary to the current approach our algorithm guarantees that the extracted core subnetwork is connected agreeing with current evidence that the core structural network is tightly connected. Second, our algorithm shows enhanced performance when used as feature selection approach for connectivity analysis on populations.

6.4.6. On the complexity of the representation of simplicial complexes by trees

Participant: Dorian Mazauric.

In collaboration with Jean-Daniel Boissonnat (DataShape team, Université Côte d'Azur, Inria).

In this paper [14], we investigate the problem of the representation of simplicial complexes by trees. We introduce and analyze local and global tree representations. We prove that the global tree representation is more efficient in terms of time complexity for searching a given simplex and we show that the local tree representation is more efficient in terms of size of the structure. The simplicial complexes are modeled by hypergraphs. We then prove that the associated combinatorial optimization problems are very difficult to solve and to approximate even if the set of maximal simplices induces a planar graph of maximum degree at most three or a bounded degree hypergraph. However, we prove polynomial time algorithms that compute constant factor approximations and optimal solutions for some classes of instances.

6.4.7. Well balanced designs for data placement

Participant: Dorian Mazauric.

In collaboration with Jean-Claude Bermond (COATI project-team, Université Côte D'Azur, Inria, I3S / CNRS), Alain Jean-Marie (MAESTRO project-team, Université Côte D'Azur, Inria) and Joseph Yu (Department of Mathematics, UFV, Abbotsford, BC, Canada).

The problem we consider in [13] is motivated by data placement, in particular data replication in distributed storage and retrieval systems. We are given a set V of v servers along with b files (data, documents). Each file is replicated on exactly k servers. A placement consists in finding a family of b subsets of V (representing the files) called blocks, each of size k. Each server has some probability to fail and we want to find a placement that minimizes the variance of the number of available files. It was conjectured that there always exists an optimal placement (with variance better than any other placement for any value of the probability of failure). We show that the conjecture is true, if there exists a well-balanced design, that is a family of blocks, each of size k, such that each j-element subset of V, $1 \leq j \leq k$, belongs to the same or almost the same number of blocks (difference at most one). The existence of well-balanced designs is a difficult problem as it contains, as a subproblem, the existence of Steiner systems. We completely solve the case math formula and give bounds and constructions for math formula and some values of v and b.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral contracts with industry

In this section, we describe the collaboration between ABS and MS Vision (http://msvision.eu/), and company based in the Netherlands. MSVision was created in 2004 and currently involves 20 employees; it is a worldwide leader in delivering tailored hardware solutions to the mass spectrometry community. As detailed below, the collaboration aims at strengthening the offer of the company on the algorithmic and software sides.

This collaboration is funded by the Instituts Carnots (http://www.instituts-carnot.eu/en).

7.1.1. Context

Protein complexes underlie most biological functions, so that studying such complexes in native conditions (intact molecular species taken in solution) is of paramount importance in biology and medicine. Unfortunately, the two leading experimental techniques to date, X ray crystallography and cryo electron microscopy, involve aggressive sample reparation (sample crystallization and sample freezing in amorphous ice, respectively) which may damage the structures and/or create artifacts. These experimental constraints legitimate the use of mass spectrometry (MS) to study biomolecules and their complexes under native conditions, using electrospray ionization (ESI), a soft ionization technique developed by John Fenn (Nobel prize in chemistry, 2002). MS actually delivers information on the masses of the molecular species studied, from which further information on the stoichiometry, topology and contacts between subunits can be inferred. Thanks to ESI, MS is expected to play a pivotal role in biology to unravel the structure of macromolecular complexes underlying all major biological processes, in medicine and biotechnology to understand the complex patterns of molecules involved in pathways, and also in biotechnologies for quality checks.

7.1.2. Specific goals

A mass spectrometer delivers a mass spectrum, i.e. an histogram representing the relative abundance of the ions (ionized proteins or protein complexes in our case), as a function of their mass-to-charge (m/z) ratio. Deconvoluting a mass spectrum means transforming it into a human readable mass histogram. Due to the nature of the ESI process (i.e. the inclusion of solvent and various other molecules) and the intrinsic variability of the studied biomolecules in native conditions, the interpretation of such spectra is delicate. Methods currently used are of heuristic nature, failing to satisfactorily handle the aforementioned difficulties. The goal of this collaboration is to develop optimal algorithms and the associated software to fill the critical gap of mass spectra deconvolution. The benefits for the analyst will be twofold, namely time savings, and the identification of previously undetected components. Upon making progress on the deconvolution problem, the collaboration will be expanded on the geometric and topological modeling of large macro-molecular assemblies, a topic to which ABS recently made significant contributions [2], [3].

8. Dissemination

8.1. Promoting scientific activities

8.1.1. Scientific Events Organisation

8.1.1.1. General Chair, Scientific Chair

Frédéric Cazals, together with C. Robert (IBPC, CNRS Paris) and J. Cortés (LAAS, CNRS Toulouse) organized the *energy landscapes* workshop, an international gathering devoted to all topics revolving around energy landscapes, as encountered in physics, chemistry, biochemistry, biology, applied mathematics, and computer science. See details at https://eland2016.inria.fr/.

8.1.2. Scientific Events Selection

8.1.2.1. Member of the Conference Program Committees

– Frédéric Cazals was member of the following program committees:

- Symposium On Geometry Processing
- Shape Modeling International: 2016
- Symposium on Solid and Physical Modeling
- Intelligent Systems for Molecular Biology (ISMB), PC member of Protein Interactions & Molecular Networks
- International Conference on Pattern Recognition in Bioinformatics

8.1.3. Journal

8.1.3.1. Reviewer - Reviewing Activities

– Frédéric Cazals reviewed papers for the following journals:

- The International Journal of Computational Geometry and Applications
- Bioinformatics
- The Journal of Immunology

8.1.4. Invited Talks and Presentations

– Frédéric Cazals gave the following invited talks:

- *Energy landscapes: sampling, analysis*, Congrès de la Société Française de Biophysique – Structural biology meets biophysics, Obernai. December 2016.
- *Modeling energy landscapes of biomolecular systems*, Ecole Normale Supérieure de Cachan. September 2016.
- *Novel structural parameters of Ig-Ag complexes yield a quantitative description of interaction specificity and binding affinity*, Structural Aspects of Infectious Disease, Cambridge, UK, August 2016.
- *Energy landscapes: sampling, analysis, and comparison*, Energy Landscapes Workshop, Porquerolles. July 2016.
- *Improved understanding of protein dynamics via energy landscape sampling, analysis, and comparison*, TSRC on protein dynamics. Les Houches, March 2016.

– Romain Tetley gave the following invited talk:

- *A bootstrap method for detecting structurally conserved motifs*, Energy Landscapes Workshop, Porquerolles. July 2016.

– Poster presentations:

- Dorian Mazauric presented the following poster:
  Unveiling Contacts within Macro-molecular Assemblies by solving Minimum Weight Connectivity Inference Problems, Congrès de la Société Française de Biophysique – Structural biology meets biophysics, Obernai. December 2016.
- Augustin Chevallier presented the following poster:
  Towards free energy calculations for biomolecules: generic Wang-Landau algorithm with automatic parameters selection, Congrès de la Société Française de Biophysique – Structural biology meets biophysics, Obernai. December 2016.
- Dorian Mazauric presented the following poster:
  Mass Transportation Problems with Connectivity Constraints and Energy Landscape Comparison, Energy Landscapes Workshop, Porquerolles. July 2016.

8.1.5. Leadership within the Scientific Community

– FrédéricCazals:

- 2010-.... Member of the steering committee of the *GDR Bioinformatique Moleculaire*, for the *Structure and macro-molecular interactions* theme.

8.1.6. Scientific Expertise

– Frédéric Cazals acted as expert for the *Italian Research and University Evaluation Agency (ANVUR)*.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Master: Frédéric Cazals (Inria ABS) and S. Oudot (Inria Saclay), *Foundations of Geometric Methods in Data Analysis*, Data Sciences Program, Department of Applied Mathematics, Ecole Centrale Paris. (http://www-sop.inria.fr/abs/teaching/centrale-FGMDA/centrale-FGMDA.html)

Master: Frédéric Cazals and Dorian Mazauric (Inria ABS), *Algorithmic problems in computational structural biology*, 24h, Master of Science in Computational Biology from the University of Nice Sophia Antipolis, France, see http://cbb.unice.fr.

8.2.2. Supervision

PhD thesis, defended, December 2016. Simon Marillet, *Modeling the antibody response: from the structure of immunoglobulins - antigen complexes to the clonal complexity of heavy chain repertoires*, University of Nice Sophia Antipolis. The thesis is co-advised by Frédéric Cazals and Pierre Boudinot (INRA Jouy-en-Josas).

PhD thesis, ongoing. Romain Tetley, *Structural alignments: beyond the rigid case*, University of Nice Sophia Antipolis. Under the supervision of Frédéric Cazals.

PhD thesis, ongoing. Augustin Chevallier, *Sampling biomolecular systems*, University of Nice Sophia Antipolis. Under the supervision of Frédéric Cazals.

Postdoctoral research of Rémi Watrigant, 2016 - 2018. Projet de Recherche Exploratoire (Inria). *Improving inference algorithms for macromolecular structure determination*. Under the supervision of Dorian Mazauric and Frédéric Havet (Inria COATI project-team).

8.2.3. Juries

– Frédéric Cazals:

- Huaxiong Ding, University of Lyon, December 2016. Committee member. *Combining 2D facial Texture and 3D face morphology for estimating people soft biometrics: gender, facial expression.* Advisors: Liming Chen and Jean-Marie Morvan.

8.3. Popularization

8.3.1. Dissemination of scientific culture

Participant: Dorian Mazauric, member of the group of Médiation et Animation des MAthématiques, des Sciences et Techniques Informatiques et des Communications (MASTIC), Inria Sophia Antipolis - Méditerranée.

8.3.1.1. Publications and ressources.

- 2016. *Graphes et Algorithmes – Jeux grandeur nature*. Dorian Mazauric, en collaboration avec Laurent Giauffret, Direction des Services Départementaux de l'Éducation Nationale (DSDEN) des Alpes-Maritimes. [https://hal.inria.fr/hal-01366804]
- 2016. *Graphes et Algorithmes - Diffusion de l'information scientifique.* Dorian Mazauric. [https://hal.inria.fr/hal-01383665]
- 2016. *Information et communication : la Théorie des Graphes*. Jean-Claude Bermond et Dorian Mazauric. Fondation la main à la pâte. To appear.

8.3.1.2. Fête de la Science en PACA.

- 22-23/10/2016. Village des sciences et de l'innovation au Palais des Congrès d'Antibes Juan-les-Pins. Fête de la Science 2016. *a) La magie des graphes et du binaire. b) Algorithmes grandeur nature. c) Pas besoin de réfléchir, les ordinateurs calculent tellement vite ? Théorie des graphes et algorithmique pour les réseaux.* [https://www.inria.fr/centre/sophia/agenda/fete-de-la-science-2016]

- 10-12/10/2016. Fête de la Science au collège Yves Montand, Vinon-sur-Verdon. Institut Esope 21. *a) La magie des graphes et du binaire. b) Algorithmes grandeur nature. c) Pas besoin de réfléchir, les ordinateurs calculent tellement vite ? Théorie des graphes et algorithmique pour les réseaux.* [https://www.inria.fr/centre/sophia/agenda/fete-de-la-science-2016]

8.3.1.3. Stage MathC2+ à Inria Sophia Antipolis - Méditerranée.

- 15-16/06/2016. Activité pour une quarantaine de lycéens des Alpes-Maritimes (accueillis à Inria Sophia Antipolis - Méditerranée durant 4 jours). *Algorithmes grandeur nature pour le calcul d'un arbre couvrant de poids minimum (application pour la conception d'un réseau électrique).* Présentation aux lycéens du stage. *Pas besoin de réfléchir, les ordinateurs calculent tellement vite ? Théorie des graphes et algorithmique pour les réseaux.* [http://www.inria.fr/centre/sophia/actualites/mathc2-40-lyceens-des-alpes-maritimes-en-immersion-au-coeur-d-un-centre-de-recherche] [https://youtu.be/rLj5I1Gu1uI]

8.3.1.4. Interventions à l'ÉSPÉ de l'Académie de Nice.

- 8-15/03/2016. Organisation d'un atelier à l'École Supérieure du Professorat et de l'Éducation (ÉSPÉ) de l'Académie de Nice (site de Stéphen Liégeard) en collaboration avec l'Inspection Académique (avec Laurent Giauffret). Animation, avec des étudiants de l'ÉSPÉ, pour 360 élèves de CM1 et de CM2. *La magie des graphes et du binaire, algorithmes et jeux (réseaux de tri).* [http://www.inria.fr/actualite/agenda/semaine-des-mathematiques-au-mois-de-mars?utm_content=buffer4b539&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer] [http://www2.ac-nice.fr/DSDEN06/cid101665/des-chercheurs-de-retour-a-l-ecole.html]

8.3.1.5. Formations pour les enseignants en collaboration avec la DSDEN des Alpes-Maritimes.

- 15-19/09/2016. Préparation avec Laurent Giauffret d'une formation pour 17 enseignants de cycle 3 (cours moyen d'enseignement élémentaire). *a) Présentation d'Inria et du dispositif ASTEP. b) Graphes et Algorithmes : théorie et mise en pratique avec des jeux. c) Présentation de Thymio et mise en avant des possibilités offertes. d) Présentation du logiciel Scratch par le Maître Assistant Informatique de circonscription.*

8.3.1.6. Conférences dans des lycées dans le cadre du dispositif régional "Science Culture".

- 26/01/2016. Conférence au lycée Amiral de Grasse (classes de seconde). *a) Pas besoin de réfléchir, les ordinateurs calculent tellement vite ? Théorie des graphes et algorithmique pour les réseaux. b) La magie des graphes et du binaire.*
- 21/01/2016. Conférence au lycée Amiral de Grasse (classes de terminale). *a) Pas besoin de réfléchir, les ordinateurs calculent tellement vite ? Théorie des graphes et algorithmique pour les réseaux. b) La magie des graphes et du binaire.*

8.3.1.7. Conférences dans des collèges des Alpes-Maritimes.

- 05/12/2016. Conférence au collège Jules Verne de Cagnes-sur-Mer (deux classe de sixième) (avec Rémi Watrigant). *La magie des graphes et du binaire.*

8.3.1.8. Conférences dans des écoles primaires des Alpes-Maritimes dans le cadre d'ASTEP.

- 22/03/2016. Conférence à l'école élémentaire de La Tournière, Antibes (classe de CE2). *La magie des graphes et du binaire, algorithmes et jeux (algorithmes grandeur nature pour trier, jeux combinatoires...).*
- 18/03/2016. Conférence à l'école élémentaire Langevin 2, Vallauris (classe de CP). *La magie des graphes et du binaire, algorithmes et jeux (algorithmes grandeur nature pour trier, jeux combinatoires...).*

8.3.1.9. Autres présentations.

- 05/01/2016. Présentation à des lycéens d'Australie et de Nouvelle-Zélande (classes de secondes) à Inria Sophia Antipolis - Méditerranée. *La magie des graphes et du binaire.*

9. Bibliography

Major publications by the team in recent years

[1] F. Cazals, P. Kornprobst (editors). *Modeling in Computational Biology and Medicine: A Multidisciplinary Endeavor*, Springer, 2013 [*DOI :* 10.1007/978-3-642-31208-3], http://hal.inria.fr/hal-00845616

[2] D. Agarwal, J. Araujo, C. Caillouet, F. Cazals, D. Coudert, S. Pérennes. *Connectivity Inference in Mass Spectrometry based Structure Determination*, in "European Symposium on Algorithms (Springer LNCS 8125)", Sophia Antipolis, France, H. Bodlaender, G. Italiano (editors), Springer, 2013, pp. 289–300, http://hal.inria.fr/hal-00849873

[3] D. Agarwal, C. Caillouet, D. Coudert, F. Cazals. *Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems*, in "Molecular and Cellular Proteomics", 2015, vol. 14, pp. 2274–2282 [*DOI :* 10.1074/mcp.M114.047779], https://hal.archives-ouvertes.fr/hal-01078378

[4] F. Cazals, F. Chazal, T. Lewiner. *Molecular shape analysis based upon the Morse-Smale complex and the Connolly function*, in "ACM SoCG", San Diego, USA, 2003, pp. 351-360

[5] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, C. Robert. *Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison*, in "J. of Computational Chemistry", 2015, vol. 36, n^{o} 16, pp. 1213–1231 [*DOI :* 10.1002/jcc.23913], https://hal.archives-ouvertes.fr/hal-01076317

[6] F. Cazals, T. Dreyfus, S. Sachdeva, N. Shah. *Greedy Geometric Algorithms for Collections of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining*, in "Computer Graphics Forum", 2014, vol. 33, n^{o} 6, pp. 1–17 [*DOI :* 10.1111/cgf.12270], http://hal.inria.fr/hal-00777892

[7] F. Cazals, J. Giesen. *Delaunay Triangulation Based Surface Reconstruction*, in "Effective Computational Geometry for curves and surfaces", J.-D. Boissonnat, M. Teillaud (editors), Springer-Verlag, Mathematics and Visualization, 2006

[8] F. Cazals, C. Karande. *An algorithm for reporting maximal c-cliques*, in "Theoretical Computer Science", 2005, vol. 349, n^{o} 3, pp. 484–490

[9] T. Dreyfus, V. Doye, F. Cazals. *Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models*, in "Proteins: structure, function, and bioinformatics", 2012, vol. 80, n^{o} 9, pp. 2125–2136

[10] T. Dreyfus, V. Doye, F. Cazals. *Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Sub-complexes*, in "Proteins: structure, function, and bioinformatics", 2013, vol. 81, n^{o} 11, pp. 2034–2044 [*DOI :* 10.1002/prot.24313], http://hal.inria.fr/hal-00849795

[11] N. Malod-Dognin, A. Bansal, F. Cazals. *Characterizing the Morphology of Protein Binding Patches*, in "Proteins: structure, function, and bioinformatics", 2012, vol. 80, n^{o} 12, pp. 2652–2665

[12] S. Marillet, P. Boudinot, F. Cazals. *High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions*, in "Proteins: structure, function, and bioinformatics", 2015, vol. 1, n^{o} 84, pp. 9–20 [*DOI :* 10.1002/prot.24946], https://hal.inria.fr/hal-01159641

Publications of the year

Articles in International Peer-Reviewed Journals

[13] J.-C. Bermond, A. Jean-Marie, D. Mazauric, J. Yu. *Well Balanced Designs for Data Placement*, in "Journal of Combinatorial Designs", February 2016, vol. 24, n^{o} 2, pp. 55-76 [*DOI :* 10.1002/jcd.21506], https://hal.inria.fr/hal-01223288

[14] J.-D. Boissonnat, D. Mazauric. *On the complexity of the representation of simplicial complexes by trees*, in "Theoretical Computer Science", February 2016, vol. 617, 17 p. [*DOI :* 10.1016/j.tcs.2015.12.034], https://hal.inria.fr/hal-01259806

[15] J. Carr, D. Mazauric, F. Cazals, D. J. Wales. *Energy landscapes and persistent minima*, in "The Journal of Chemical Physics", February 2016, vol. 144, n^{o} 5 [*DOI :* 10.1063/1.4941052], https://hal.inria.fr/hal-01423280

[16] D. Coudert, D. Mazauric, N. Nisse. *Experimental Evaluation of a Branch and Bound Algorithm for Computing Pathwidth and Directed Pathwidth*, in "ACM Journal of Experimental Algorithmics", 2016, vol. 21, n^{o} 1, 23 p. [*DOI :* 10.1145/2851494], https://hal.inria.fr/hal-01266496

[17] S. Grudinin, M. Kadukova, A. Eisenbarth, S. Marillet, F. Cazals. *Predicting binding poses and affinities for protein-ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation*, in "Journal of Computer-Aided Molecular Design", September 2016, vol. 30, n^{o} 9, pp. 791–804 [*DOI :* 10.1007/s10822-016-9976-2], https://hal.inria.fr/hal-01377738

[18] A. Roth, T. Dreyfus, C. H. Robert, F. Cazals. *Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes*, in "Journal of Computational Chemistry", January 2016, vol. 37, n^{o} 8, 14 p. , https://hal.inria.fr/hal-01423282

International Conferences with Proceedings

[19] D. Wassermann, D. Mazauric, G. Gallardo-Diez, R. Deriche. *Extracting the Core Structural Connectivity Network: Guaranteeing Network Connectedness Through a Graph-Theoretical Approach*, in "MICCAI 2016", Athens, Greece, September 2016, https://hal.inria.fr/hal-01333301

Research Reports

[20] F. Cazals, T. Dreyfus. *The Structural Bioinformatics Library: modeling in biomolecular science and beyond*, Inria, October 2016, n^{o} RR-8957, https://hal.inria.fr/hal-01379635

[21] F. Cazals, D. Mazauric. *Optimal transportation problems with connectivity constraints*, Inria Sophia Antipolis ; Université Côte d'Azur, December 2016, n^{o} RR-8991, 24 p. , https://hal.inria.fr/hal-01411117

[22] F. Cazals, D. Mazauric, R. Tetley. *Clustering stability revealed by matchings between clusters of clusters*, Inria Sophia Antipolis ; Université Côte d'Azur, December 2016, n^{o} RR-8992, 51 p. , https://hal.inria.fr/hal-01410396

[23] S. Marillet, M.-P. Lefranc, P. Boudinot, F. Cazals. *Novel structural parameters of Ig -Ag complexes yield a quantitative description of interaction specificity and binding affinity*, Inria Sophia Antipolis, October 2016, n^{o} RR-8963, https://hal.inria.fr/hal-01381795

Scientific Popularization

[24] D. Mazauric. *Graphes et Algorithmes – Jeux grandeur nature*, Inria, 2016, 401 p. , https://hal.inria.fr/hal-01366804

[25] D. Mazauric. *Graphes et Algorithmes - Diffusion de l'information scientifique*, 2016, 403 p. , Slides de médiation scientifique pour comprendre les graphes et les algorithmes de manière ludique, https://hal.inria.fr/hal-01383665

References in notes

[26] F. Alber, S. Dokudovskaya, L. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B. Chait, M. Rout, A. Sali. *Determining the architectures of macromolecular assemblies*, in "Nature", Nov 2007, vol. 450, pp. 683-694

[27] F. Alber, S. Dokudovskaya, L. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B. Chait, A. Sali, M. Rout. *The molecular architecture of the nuclear pore complex*, in "Nature", 2007, vol. 450, n^{o} 7170, pp. 695–701

[28] F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali. *Integrating Diverse Data for Structure Determination of Macromolecular Assemblies*, in "Ann. Rev. Biochem.", 2008, vol. 77, pp. 11.1–11.35

[29] O. Becker, A. D. Mackerell, B. Roux, M. Watanabe. *Computational Biochemistry and Biophysics*, M. Dekker, 2001

[30] A.-C. Camproux, R. Gautier, P. Tuffery. *A Hidden Markov Model derived structural alphabet for proteins*, in "J. Mol. Biol.", 2004, pp. 591-605

[31] M. L. Connolly. *Analytical molecular surface calculation*, in "J. Appl. Crystallogr.", 1983, vol. 16, n^{o} 5, pp. 548–558

[32] R. Dunbrack. *Rotamer libraries in the 21st century*, in "Curr Opin Struct Biol", 2002, vol. 12, n^{o} 4, pp. 431-440

[33] A. Fernandez, R. Berry. *Extent of Hydrogen-Bond Protection in Folded Proteins: A Constraint on Packing Architectures*, in "Biophysical Journal", 2002, vol. 83, pp. 2475-2481

[34] A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, Freeman, 1999

[35] M. Gerstein, F. Richards. *Protein geometry: volumes, areas, and distances*, in "The international tables for crystallography (Vol F, Chap. 22)", M. G. Rossmann, E. Arnold (editors), Springer, 2001, pp. 531–539

[36] H. Gohlke, G. Klebe. *Statistical potentials and scoring functions applied to protein-ligand binding*, in "Curr. Op. Struct. Biol.", 2001, vol. 11, pp. 231-235

[37] J. JANIN, S. WODAK, M. LEVITT, B. MAIGRET. *Conformations of amino acid side chains in proteins*, in "J. Mol. Biol.", 1978, vol. 125, pp. 357–386

[38] V. K. KRIVOV, M. KARPLUS. *Hidden complexity of free energy surfaces for peptide (protein) folding*, in "PNAS", 2004, vol. 101, n^{o} 41, pp. 14766-14770

[39] E. MEERBACH, C. SCHUTTE, I. HORENKO, B. SCHMIDT. *Metastable Conformational Structure and Dynamics: Peptides between Gas Phase and Aqueous Solution*, in "Analysis and Control of Ultrafast Photoinduced Reactions. Series in Chemical Physics 87", O. KUHN, L. WUDSTE (editors), Springer, 2007

[40] I. MIHALEK, O. LICHTARGE. *On Itinerant Water Molecules and Detectability of Protein-Protein Interfaces through Comparative Analysis of Homologues*, in "JMB", 2007, vol. 369, n^{o} 2, pp. 584–595

[41] J. MINTSERIS, B. PIERCE, K. WIEHE, R. ANDERSON, R. CHEN, Z. WENG. *Integrating statistical pair potentials into protein complex prediction*, in "Proteins", 2007, vol. 69, pp. 511–520

[42] M. PETTINI. *Geometry and Topology in Hamiltonian Dynamics and Statistical Mechanics*, Springer, 2007

[43] E. PLAKU, H. STAMATI, C. CLEMENTI, L. KAVRAKI. *Fast and Reliable Analysis of Molecular Motion Using Proximity Relations and Dimensionality Reduction*, in "Proteins: Structure, Function, and Bioinformatics", 2007, vol. 67, n^{o} 4, pp. 897–907

[44] D. RAJAMANI, S. THIEL, S. VAJDA, C. CAMACHO. *Anchor residues in protein-protein interactions*, in "PNAS", 2004, vol. 101, n^{o} 31, pp. 11287-11292

[45] D. REICHMANN, O. RAHAT, S. ALBECK, R. MEGED, O. DYM, G. SCHREIBER. *From The Cover: The modular architecture of protein-protein binding interfaces*, in "PNAS", 2005, vol. 102, n^{o} 1, pp. 57-62

[46] F. RICHARDS. *Areas, volumes, packing and protein structure*, in "Ann. Rev. Biophys. Bioeng.", 1977, vol. 6, pp. 151-176

[47] G. RYLANCE, R. JOHNSTON, Y. MATSUNAGA, C.-B. LI, A. BABA, T. KOMATSUZAKI. *Topographical complexity of multidimensional energy landscapes*, in "PNAS", 2006, vol. 103, n^{o} 49, pp. 18551-18555

[48] G. SCHREIBER, L. SERRANO. *Folding and binding: an extended family business*, in "Current Opinion in Structural Biology", 2005, vol. 15, n^{o} 1, pp. 1–3

[49] M. SIPPL. *Calculation of Conformational Ensembles from Potential of Mean Force: An Approach to the Knowledge-based prediction of Local Structures in Globular Proteins*, in "J. Mol. Biol.", 1990, vol. 213, pp. 859-883

[50] C. SUMMA, M. LEVITT, W. DEGRADO. *An atomic environment potential for use in protein structure prediction*, in "JMB", 2005, vol. 352, n^{o} 4, pp. 986–1001

[51] S. WODAK, J. JANIN. *Structural basis of macromolecular recognition*, in "Adv. in protein chemistry", 2002, vol. 61, pp. 9–73