Activity Report 2014

Project-Team ABS

Algorithms, Biology, Structure

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Computational Biology

Table of contents

Project-Team ABS

Keywords: Computational Structural Biology, Protein-protein Interactions, Protein Assemblies, Computational Geometry, Computational Topology

Creation of the Project-Team: 2008 July 01.

1. Members

Research Scientist

Frédéric Cazals [Team leader, Inria, Senior Researcher, HdR]

Engineer

Tom Dreyfus [Inria]

PhD Students

Deepesh Agarwal [Inria, until September 2014]
Alix Lhéritier [Inria]
Simon Marillet [INRA]
Christine Roth [Inria, until September 2014]
Romain Tetley [University of Nice Sophia Antipolis, since November 2014]

Administrative Assistant

Florence Barbara [Inria]

Others

Charles Robert [CNRS, External Collaborator, HdR]
Darsh Shah [Inria, intern from ITT Bombay,from May 2014 until July 2014]
Romain Tetley [Inria, intern from University of Nice Sophia Antipolis, from March 2014 until August 2014]

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 [39]. 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 [37], [26] and later Connolly [22], 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) [28], 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.

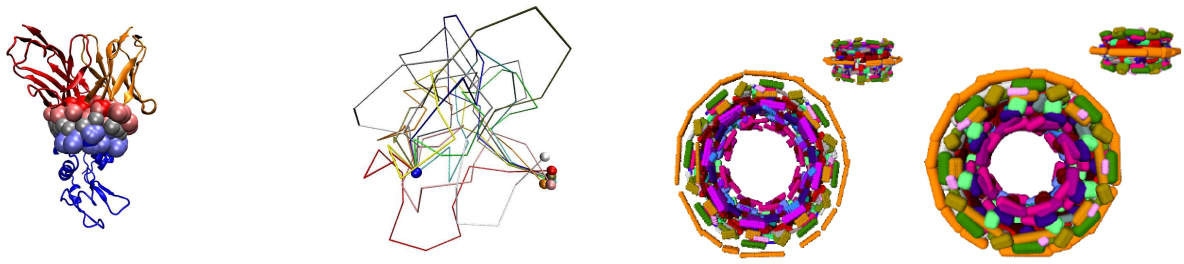
3. Research Program

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.



(a) (b) (c)

Figure 1. ***Geometric constructions in computational structural biology.*** *(a) An antibody-antigen complex, with interface atoms identified by our Voronoi based interface model [8], [1]. This model is instrumental in mining correlations between structural and biological as well as biophysical properties of protein complexes. (b) A diverse set of conformations of a backbone loop, selected thanks to a geometric optimization algorithm [10]. 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.*

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 [39]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [42]. Current investigations follow two routes. From the experimental perspective [25], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [36]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [31].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change [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 [20], the usual way to design scoring functions is to mimic the so-called potentials of mean force. To put it briefly, one reverts Boltzmann's law, that is, denoting $p_i(r)$ the probability of two atoms –defining type i– to be located at distance r, the (free) energy assigned to the pair is computed as $E_i(r) = -kT \log p_i(r)$. Estimating from the PDB one function $p_i(r)$ for each type of pair of atoms, the energy of a complex is computed as the sum of the energies of the pairs located within a distance threshold [40], [27]. To compare the energy thus obtained to a reference state, one may compute $E = \sum_i p_i \log p_i/q_i$, with p_i the observed frequencies, and q_i the frequencies stemming from an a priori model [32]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

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

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

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 [8]. To elaborate at the atomic level, one seeks a structural alphabet encoding the spatial structure of proteins. At the side-chain and backbone level, an example of such alphabet is that of [21]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [41], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond – a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [24].

In the dynamic setting, one wishes to understand whether selected (hot) residues exhibit specific dynamic properties, so as to serve as anchors in a binding process [35]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

3.3. Modeling Macro-molecular Assemblies

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

3.3.1. Reconstruction by Data Integration

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [19]. In a nutshell, the strategy is reminiscent from NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This strategy has been in particular used to propose plausible models of the Nuclear Pore Complex [18], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

3.3.2. Modeling with Uncertainties and Model Assessment

Reconstruction by data integration requires three ingredients. First, a parametrized model must be adopted, typically a collection of balls to model a protein with pseudo-atoms. Second, as in NMR, a functional measuring the agreement between a model and the data must be chosen. In [17], this functional is based upon *restraints*, namely penalties associated to the experimental data. Third, an optimization scheme must be selected. The design of restraints is notoriously challenging, due to the ambiguous nature and/or the noise level of the data. For example, Tandem Affinity Purification (TAP) gives access to a *pullout* i.e. a list of protein types which are known to interact with one tagged protein type, but no information on the number of complexes or on the stoichiometry of proteins types within a complex is provided. In cryo-EM, the envelope enclosing an assembly is often imprecisely defined, in particular in regions of low density. For immuno-EM labelling experiments, positional uncertainties arise from the microscope resolution.

These uncertainties coupled with the complexity of the functional being optimized, which in general is non convex, have two consequences. First, it is impossible to single out a unique reconstruction, and a set of plausible reconstructions must be considered. As an example, 1000 plausible models of the NPC were reported in [17]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual

models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

3.4. Modeling the Flexibility of Macro-molecules

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

Proteins in vivo vibrate at various frequencies: high frequencies correspond to small amplitude deformations of chemical bonds, while low frequencies characterize more global deformations. This flexibility contributes to the entropy thus the `free energy` of the system *protein - solvent*. From the experimental standpoint, NMR studies generate ensembles of conformations, called `conformers`, and so do molecular dynamics (MD) simulations. Of particular interest while investigating flexibility is the notion of correlated motion. Intuitively, when a protein is folded, all atomic movements must be correlated, a constraint which gets alleviated when the protein unfolds since the steric constraints get relaxed [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 [23]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [38]. The standard methodology to analyze such a point cloud consists of resorting to normal modes. Recently, though, more elaborate methods resorting to more local analysis [34], to Morse theory [29] and to analysis of meta-stable states of time series [30] have been proposed.

3.5. Algorithmic Foundations

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

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

3.5.1. Modeling Interfaces and Contacts

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

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

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

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

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 [33].

From an algorithmic standpoint, such questions are reminiscent of *shape learning*. Given a collection of samples on an (unknown) *model*, *learning* consists of guessing the model from the samples – the result of this process may be called the *reconstruction*. In doing so, two types of guarantees are sought: topologically speaking, the reconstruction and the model should (ideally!) be isotopic; geometrically speaking, their Hausdorff distance should be small. Motivated by applications in Computer Aided Geometric Design, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years [5]. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison; the study of Morse-like constructions stemming from distance functions to points; the analysis of topological invariants of the model and the samples, and their comparison.

4. New Software and Platforms

4.1. Software

Until October 2014, ABS was distributing isolated programs to solve selected tasks in computational structural biology, including:

- `vorpatch` and `compatch`: Modeling and Comparing Protein Binding Patches,
- `intervor`: Modeling Macro-molecular Interfaces,
- `vorlume`: Computing Molecular Surfaces and Volumes with Certificates,
- `ESBTL`: the Easy Structural Biology Template Library.

This software has been completely repackaged within the *Structural Bioinformatics Library*, a C++ library developed in the scope of an Inria supported *ADT*. The SBL will be released early 2015. Below, we briefly review its spirit and contents.

The Structural Bioinformatics Library (SBL): overview. The Structural Bioinformatics Library (SBL) is a generic C++/python library providing combinatorial, geometric and topological tools to solve problems in computational structural biology (CSB). Its design is meant to accommodate both the variety of models coding the physical and chemical properties of macro-molecular systems, and the variety of operations undertaken on these models. The models supported either consist of unions of balls (van der Waals models, solvent accessible models), or representations of conformations based on Cartesian or internal coordinates (distances and angles between the atoms). The operations provided revolve around the problem of understanding the relationship between the structure and the function of macro-molecules and their complexes, and deal with complementary aspects, namely geometric, topological, and combinatorial methods are used to foster our understanding of bio-physical and biological properties. Software development in this context is especially challenging due to the interactions between these complex models and operations.

To accommodate this complexity, software components of the SBL are organized into four categories:

- SBL-APPLICATIONS**:** end-user 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.

The SBL for end-users. End users will find in the SBL portable applications running on Linux, and MacOS. These applications split into the following categories:

- **Space Filling Models:** applications dealing with molecular models defined by unions of balls. Current statistics are:
  - # classes: 151
  - # lines of C++/python: 65,000
  - # pages of documentation (user + reference manuals): $\sim$ 1000
- **Conformational Analysis:** applications dealing with molecular flexibility. Current statistics are:
  - # classes: 110
  - # lines of C++/python: 49,000
  - # pages of documentation (user + reference manuals): $\sim$ 800
- **Data Analysis:** 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.
- **Large assemblies:** applications dealing with macro-molecular assemblies involving from tens to hundreds of macro-molecules.

The SBL for developers. Development with the SBL may occur at two levels.

Low level developments may use classes from 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.

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

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.

Intermediate level developments should be based upon modules, since 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.

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 (e.g. molecular dynamics 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/).

Releases, distribution, and licence. The SBL will be released under a proprietary open source licence. In a nutshell, academic users can use and modify the code at their discretion, for private purposes. But distributing these changes, or doing business with the SBL is forbidden. However, novel capabilities matching the design choices of the library will be welcome, and may be integrated.

The source code will be distributed from http://structural-bioinformatics-library.org/, as a tarball and also via a git repository. Bugzilla will be used to handle user's feedback and bug tracking.

The releases are scheduled as follows:

- February 2015: applications from the *space filling model* group, and the accompanying low level classes.
- April 2015: applications from *conformational analysis* group, and the accompanying low level classes.
- July 2015: applications from *large assemblies* group, and the accompanying low level classes.

5. New Results

5.1. Highlights of the Year

In 2014, two achievements are worth noticing:
Analysis of large assemblies using native mass spectrometry data. Native mass spectrometry is about to revolutionize structural biology, since such experiments give access to the composition in terms of subunits of large macro-molecular assemblies, usually beyond reach for classical experimental techniques. In this context, we designed an algorithm to infer pairwise contacts within subunits of large macro-molecular assemblies – see section 5.3.1. To the best of our knowledge, our algorithm is the only one whose performances can be precisely analyzed, the contenders being of heuristic nature.

Analysis and comparison of conformational ensembles and sampled energy landscapes. A key property governing the behavior of many biophysical systems is the classical enthalpy - entropy balance, which is the root of thermodynamics. Therefore, studying the way a protein folds or the way two proteins assemble requires unveiling properties of ensembles of conformations of the system scrutinized. In this context, we designed novel methods to analyze and compare collections of conformations and the associated energy landscape – see section 5.4.1. The algorithms are based on state-of-the-art techniques from computational topology (Morse theory, Morse homology), and optimal transportation.

5.2. Modeling Interfaces and Contacts

Docking, scoring, interfaces, protein complexes, Voronoi diagrams, arrangements of balls.

The work undertaken in this vein in 2014 will be finalized in 2015.

5.3. Modeling Macro-molecular Assemblies

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

5.3.1. Connectivity Inference in Mass Spectrometry based Structure Determination

Participants: Frédéric Cazals, Deepesh Agarwal.

In collaboration with C. Caillouet, and D. Coudert, from the COATI project-team (Inria - I3S (CNRS, University of Nice Sophia Antipolis)).

Consider a set of oligomers listing the subunits involved in sub-complexes of a macro-molecular assembly, obtained e.g. using native mass spectrometry or affinity purification. Given these oligomers, connectivity inference (CI) consists of finding the most plausible contacts between these subunits, and minimum connectivity inference (MCI) is the variant consisting of finding a set of contacts of smallest cardinality. MCI problems avoid speculating on the total number of contacts, but yield a subset of all contacts and do not allow exploiting a priori information on the likelihood of individual contacts.

In this paper [14], we present two novel algorithms, MILP-W and MILP-WB. The former solves the minimum weight connectivity inference (MWCI), an optimization problem whose criterion mixes the number of contacts and their likelihood. The latter uses the former in a bootstrap fashion, to improve the sensitivity and the specificity of solution sets.

Experiments on three systems (yeast exosome, yeast proteasome lid, human eiF3), for which reference contacts are known (crystal structure, cryo electron microscopy, cross-linking), show that our algorithms predict contacts with high specificity and sensitivity, yielding a very significant improvement over previous work, typically a twofold increase in sensitivity.

The software accompanying this paper is made available, and should prove of ubiquitous interest whenever connectivity inference from oligomers is faced.

5.4. Modeling the Flexibility of Macro-molecules

Protein, flexibility, collective coordinate, conformational sampling dimensionality reduction.

5.4.1. Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison

Participants: Frédéric Cazals, Tom Dreyfus, Christine Roth.

In collaboration with D. Mazauric (Inria Sophia Antipolis Méditerranée, Geometrica) and C. Robert (IBPC / CNRS, Paris).

In this work, we present novel algorithms and software addressing four core problems in computational structural biology, namely analyzing a conformational ensemble, comparing two conformational ensembles, analyzing a sampled energy landscape, and comparing two sampled energy landscapes [15]. Using recent developments in computational topology, graph theory, and combinatorial optimization, we make two notable contributions. First, we present a generic algorithm analyzing height fields. We then use this algorithm to perform density based clustering of conformations, and to analyze a sampled energy landscape in terms of basins and transitions between them. In both cases, topological persistence is used to manage (geometric) frustration. Second, we introduce two algorithms to compare transition graphs. The first is the classical *earth mover distance* metric which depends only on local minimum energy configurations along with their statistical weights, while the second incorporates topological constraints inherent to conformational transitions.

Illustrations are provided on a simplified protein model (BLN69), whose frustrated potential energy landscape has been thoroughly studied.

The software implementing our tools is also made available, and should prove valuable wherever conformational ensembles and energy landscapes are used.

5.5. Algorithmic Foundations

Computational geometry, Computational topology, Voronoi diagrams, α-shapes, Morse theory.

5.5.1. Mass Transportation Problems with Connectivity Constraints

Participant: Frédéric Cazals.

In collaboration with D. Mazauric (Inria Sophia Antipolis Méditerranée, Geometrica).

Given two graphs, the supply and the demand graphs, we analyze the mass transportation problem between their vertices, under connectivity constraints [16]. More precisely, for every subset of supply nodes inducing a connected component of the supply graph, we require that the set of demand nodes receiving non-zero flow from this subset induces a connected component of the demand graph. As opposed to the classical problem, a.k.a the earth mover distance (EMD), which is amenable to linear programming (LP), this new problem is very difficult to solve, and we make four contributions. First, we formally introduce two optimal transportation problems, namely minimum-cost flow under connectivity constraints problem (EMD-CC) and maximum-flow under cost and connectivity constraints problem (EMD-CCC). Second, we prove that the decision version of EMD-CC is NP-complete even for very simple classes of instances. We deduce that the decision version of EMD-CCC is NP-complete, and also prove that EMD-CC is not in APX even for simple classes of instances. Third, 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. Finally, on the experimental side, we compare the transport plans computed by our greedy method against those produced by the aforementioned LP. Using synthetic landscapes (Voronoi landscapes), we show that our greedy algorithm is effective for graphs involving up to 1000 nodes. We also show the relevance of our algorithms to compare energy landscapes of biophysical systems (protein models).

5.5.2. Ciruvis: a web-based Tool for Rule Networks and Interaction Detection using Rule-based Classifiers

Participant: Simon Marillet.

In collaboration with J. Komorowski and S. Bornelöv (Uppsala University).

The use of classification algorithms is becoming increasingly important for the field of computational biology. However, not only the quality of the classification, but also its biological interpretation is important. This interpretation may be eased if interacting elements can be identified and visualized, something that requires appropriate tools and methods.

We developed a new approach to detecting interactions in complex systems based on classification [12]. Using rule-based classifiers, we previously proposed a rule network visualization strategy that may be applied as a heuristic for finding interactions. We now complement this work with Ciruvis, a web-based tool for the construction of rule networks from classifiers made of IF-THEN rules. Simulated and biological data served as an illustration of how the tool may be used to visualize and interpret classifiers. Furthermore, we used the rule networks to identify feature interactions, compared them to alternative methods, and computationally validated the findings. Rule networks enable a fast method for model visualization and provide an exploratory heuristic to interaction detection. The tool is made freely available on the web and may thus be used to aid and improve rule-based classification.

6. Partnerships and Cooperations

6.1. National Initiatives

6.1.1. Projets Exploratoires Pluridisciplinaires from CNRS/Inria/INSERM

Title: Novel approaches to characterizing flexible macromolecular systems in biology

Modeling Large Protein Assemblies with Toleranced Models

Type: Projet Exploratoire Pluri-disciplinaire (PEPS) CNRS / Inria / INSERM

Duration: one year

Coordinator: C. Robert (IBPC / CNRS)

Other partner(s): F. Cazals (Inria Sophia Antipolis Méditerranée)

Abstract: A central problem in structural biology consists of modeling the dynamics and thermodynamics of macro-molecular assemblies involving a large number of atoms (thousands to hundreds of thousands). This requires understanding the structure of the potential and free energy landscapes (PEL and FEL) of the system. A number of approaches have been developed from the physical perspective, in particular to sample the PEL of the systems scrutinized (molecular dynamics, Monte Carlo based methods). The goal of this project is orthogonal, since our aim is to enhance the processing of samplings generated by the aforementioned approaches. Our methods aim at analyzing and comparing sampled PEL and FEL, using novel methods from computational geometry, computational topology, and optimization. These methods should foster our understanding of the behavior of macro-molecular assemblies, and in the long run, they should also trigger the development of more efficient sampling algorithms.

6.2. International Initiatives

6.2.1. Participation In other International Programs

F. Cazals (Inria ABS), I. Emiris (Prof., Univ. of Athens) and S. Theodoridis (Prof., Univ. of Athens) collaborate in the scope of an Inria COLOR entitled *Discriminating and classifying in high-dimensional spaces*.

The scientific goal was to study methods and algorithms in high dimensional spaces, revolving around three problems: approximate nearest neighbors, polytope volume approximations, and classification - discrimination in high high-dimensional Spaces.

The long-term plan is to examine whether the work done so far can be combined with work by other European teams targeting a European research proposal. F. Cazals and I. Emiris participate in a FET-Open STREP proposal, entitled *Exploring the Geometry of Data*, including high-dimensional geometry, machine learning, and statisctical methods. More precisely, the collaborations proposed between the two groups bootstraps on the achievements of the COLOR, as they aim at exploring (i) incremental nearest neighbor methods in metric spaces, (ii) sampling methods for polytope volume approximation and high-dimensional space exploration, and (iii) applications in biophysics (protein docking and energy landscape exploration).

6.3. International Research Visitors

6.3.1. Visits of International Scientists

- Fasseli Coulibaly, Monash University, September 2014.

6.3.1.1. Internships

- R. Tetley, from the MSc program *Computational biology and biomedicine* from the Univ. of Nice, completed his MSc internship under the guidance of F. Cazals, on the topic *Bootstrap algorithms for structural alignments, with applications in structural virology*. Romain is now following-up as a PhD student.

- D. Shah, second year student from the IIT Bombay, completed a summer internship on the topic *Improving scoring functions for protein docking*.

7. Dissemination

7.1. Promoting Scientific Activities

7.1.1. Scientific Events Organisation

7.1.1.1. general chair, scientific chair

- F. Cazals co-organized, with C. Robert (IBPC/CNRS, Paris) and J. Cortés (LAAS/CNRS, Toulouse), the Winter School *Modeling Large Macromolecular Assemblies*, held in Sophia-Antipolis on 8-12 December 2014. The school featured lectures by Gerhard Hummer (Max Plack Institut fur Biophysik / Theoretical Biophysics, Frankfurt), Dmitri Svergun (European Molecular Biology Laboratory, Hamburg), Haim Wolfson (Tel Aviv University), and Riccardo Pellarin (UCSF / Institut Pasteur Paris), and FrédéricCazals. The event gathered 40 students from all over the world. See details at http://algosb2014.inria.fr/.

7.1.2. Scientific Events Selection

7.1.2.1. member of the conference program committee

F. Cazals was member of the following program committees:

- Symposium On Geometry Processing
- ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics

7.2. Teaching - Supervision - Juries

7.2.1. Teaching

Master: F. Cazals (Inria Sophia Antipolis Méditerranée) 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: F. Cazals, *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.

7.2.2. Supervision

(PhD thesis, ongoing) C. Roth, *Modeling the flexibility of macro-molecules: theory and applications*, University of Nice Sophia Antipolis. Advisor: F. Cazals.

(PhD thesis, ongoing) A. Lheritier, *Scoring and discriminating in high-dimensional spaces: a geometric based approach of statistical tests*, University of Nice Sophia Antipolis. Advisor: F. Cazals.

(PhD thesis, ongoing) D. Agarwal, *Towards nano-molecular design: advanced algorithms for modeling large protein assemblies*, University of Nice Sophia Antipolis. Advisor: F. Cazals.

(PhD thesis, ongoing) S. Marillet, *Modeling antibody - antigen complexes*, University of Nice Sophia Antipolis. The thesis is co-advised by F. Cazals and P. Boudinot (INRA Jouy-en-Josas).

(PhD thesis, ongoing) R. Tetley, *Structural alignments: beyond the rigid case*, University of Nice Sophia Antipolis.

7.2.3. Juries

- Didier Devaurs, University of Toulouse, October 2014. Rapporteur on the PhD thesis *Extensions of Sampling-based Approaches to Path Planning in Complex Cost Spaces: Applications to Robotics and Structural Biology*. Advisor: Juan Cortés
- Juan Cortés, University of Toulouse, April 2014. Rapporteur on the Habilitation thesis *Algorithmics of motion: from robotics, through structural biology, towards atomic-scale CAD*.

8. Bibliography

Major publications by the team in recent years

[1] B. BOUVIER, R. GRUNBERG, M. NILGES, F. CAZALS. *Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics and composition*, in "Proteins: structure, function, and bioinformatics", 2009, vol. 76, n^{o} 3, pp. 677–692

[2] F. CAZALS. *Effective nearest neighbors searching on the hyper-cube, with applications to molecular clustering*, in "Proc. 14th Annu. ACM Sympos. Comput. Geom.", 1998, pp. 222–230

[3] 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

[4] F. CAZALS, T. DREYFUS. *Multi-scale Geometric Modeling of Ambiguous Shapes with Toleranced Balls and Compoundly Weighted α-shapes*, in "Symposium on Geometry Processing", Lyon, B. LEVY, O. SORKINE (editors), 2010, pp. 1713-1722, Also as Inria Tech report 7306

[5] 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

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

[7] F. CAZALS, S. LORIOT. *Computing the exact arrangement of circles on a sphere, with applications in structural biology*, in "Computational Geometry: Theory and Applications", 2009, vol. 42, n^{o} 6-7, pp. 551–565, Preliminary version as Inria Tech report 6049

[8] F. CAZALS, F. PROUST, R. BAHADUR, J. JANIN. *Revisiting the Voronoi description of Protein-Protein interfaces*, in "Protein Science", 2006, vol. 15, n^{o} 9, pp. 2082–2092

[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] S. LORIOT, S. SACHDEVA, K. BASTARD, C. PREVOST, F. CAZALS. *On the Characterization and Selection of Diverse Conformational Ensembles*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2011, vol. 8, n^{o} 2, pp. 487–498

[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

Publications of the year

Articles in International Peer-Reviewed Journals

[12] S. BORNELÖV, S. MARILLET, J. KOMOROWSKI. *Ciruvis: a web-based tool for rule networks and interaction detection using rule-based classifiers*, in "BMC Bioinformatics", 2014, vol. 15, n° 1, 139 p. , https://hal.inria.fr/hal-00995110

[13] F. CAZALS, T. DREYFUS, S. SACHDEVA, N. SHAH. *Greedy Geometric Algorithms for Collection of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining*, in "Computer Graphics Forum", 2014, vol. 33, pp. 1 - 17 [*DOI :* 10.1111/CGF.12270], https://hal.inria.fr/hal-01110229

Research Reports

[14] D. AGARWAL, C. CAILLOUET, D. COUDERT, F. CAZALS. *Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems*, Inria, October 2014, n° RR-8622, https://hal.archives-ouvertes.fr/hal-01078378

[15] F. CAZALS, T. DREYFUS, D. MAZAURIC, A. ROTH, C. ROBERT. *Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison*, Inria, October 2014, n° RR-8610, https://hal.archives-ouvertes.fr/hal-01076317

[16] F. CAZALS, D. MAZAURIC. *Mass Transportation Problems with Connectivity Constraints, with Applications to Energy Landscape Comparison*, Inria Sophia Antipolis, October 2014, n° RR-8611, https://hal.archives-ouvertes.fr/hal-01090705

References in notes

[17] 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

[18] 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° 7170, pp. 695–701

[19] 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

[20] O. BECKER, A. D. MACKERELL, B. ROUX, M. WATANABE. *Computational Biochemistry and Biophysics*, M. Dekker, 2001

[21] A.-C. CAMPROUX, R. GAUTIER, P. TUFFERY. *A Hidden Markov Model derived structural alphabet for proteins*, in "J. Mol. Biol.", 2004, pp. 591-605

[22] M. L. CONNOLLY. *Analytical molecular surface calculation*, in "J. Appl. Crystallogr.", 1983, vol. 16, n° 5, pp. 548–558

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

[24] 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

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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

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

[34] 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

[35] 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

[36] 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

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

[38] 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

[39] 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

[40] 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

[41] 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

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