Activity Report 2015

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
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]

Engineer

Tom Dreyfus [Inria, until November 2015]

PhD Students

Deepesh Agarwal [Inria, Until May 2015]
Alix Lhéritier [Inria, Until September 2015]
Christine Roth
Simon Marillet [INRA]
Romain Tetley [Univ. Nice]
Augustin Chevallier [Univ. Nice, from October 2015]

Administrative Assistant

Florence Barbara [Inria]

Others

Nathalie Gayraud [Inria, February - August 2015]
Charles Robert [CNRS, HdR]

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 [45]. 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.

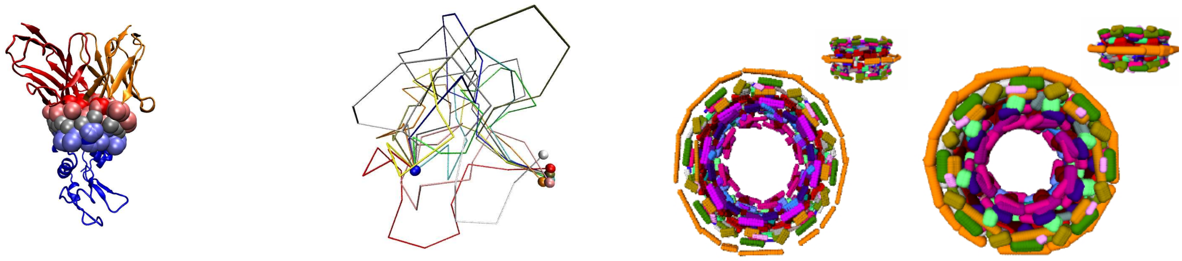
Modeling in Computational Structural Biology. Acquiring structural data is not always possible: NMR is restricted to relatively small molecules; membrane proteins do not crystallize, etc. As a matter of fact, the order of magnitude of the number of genomes sequenced is of the order of one thousand, which 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 [43], [32] and later Connolly [28], 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) [34], 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.



(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 [3]. 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.*

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 [45]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [48]. Current investigations follow two routes. From the experimental perspective [31], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [42]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [37].

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 [26], 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 [46], [33]. 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 [38]. 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 [3]. 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 [27]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [47], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond – a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [30].

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 [41]. 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

[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.

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 [25]. 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 [24], 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 [23], 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 [23]. 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.

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

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

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [44]. 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 [40], to Morse theory [35] and to analysis of meta-stable states of time series [36] have been proposed.

3.5. Algorithmic Foundations

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

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

3.5.1. Modeling Interfaces and Contacts

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

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

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

3.5.2. Modeling Macro-molecular Assemblies

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

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 [39].

From an algorithmic standpoint, such questions are reminiscent of *shape learning*. Given a collection of samples on an (unknown) *model*, *learning* consists of guessing the model from the samples – the result of this process may be called the *reconstruction*. In doing so, two types of guarantees are sought: topologically speaking, the reconstruction and the model should (ideally!) be isotopic; geometrically speaking, their Hausdorff distance should be small. Motivated by applications in Computer Aided Geometric Design, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years [6]. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison; 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 2015, several achievements are worth noticing in three realms, namely in computer science, computational structural biology, and software.

4.1.1. Computer Science

► Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces
Reference: [17]

In a nutshell: A classical problem in statistics is to decide whether two populations exhibit a statistically significant difference—the so-called two-sample test problem (TST). If so, another classical problem is to assess the magnitude of the difference—the so-called effect size calculation. While various effect size calculations were available for univariate data, hardly any existed for multivariate data.

Assessment: In this work, we provide one of the very first (if not the first) effect size calculation for multivariate data. The method combines techniques from machine learning (regression) and computational topology (topological persistence).

4.1.2. Computational Structural Biology

► High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions
Reference: [20]

In a nutshell: The binding affinity of two proteins forming a complex is a key quantity, whose estimation from structural data has remained elusive, a difficulty owing to the variety of protein binding modes. In this work, we present sparse models using up to five variables describing enthalpic and entropic variations upon binding, and a (cross-validation based) model selection procedure identifying the best sparse models built from a subset of these variables.

Assessment: Our estimation method ranks amongst the top two or three known so far, and is possibly the most accurate when applied to high resolution crystal structures. One of its key limitations (similar to contenders) is that the crystal structures of the partners and that of the complex are required. This limitation motivates our work on energy landscapes, see below.

► Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems
Reference: [14]

In a nutshell: Following the 2002 Nobel prize in chemistry of Fenn and Tanaka, and the recent developments led in particular by Carol Robinson (Oxford), native mass spectrometry is about to become a technique of major importance in structural biology, providing information on large assemblies (more than 10 subunits) studied in solution. One key question is to infer pairwise contacts between subunits from native mass spectrometry data.

Assessment: In this work, we provide a method to predict pairwise contacts between subunits of a large assembly, based on the composition of oligomers. The method is based on a mixed linear integer program, and essentially doubles the prediction performances of the method developed by Robinson et al.

► **Hybridizing Rapidly Growing Random Trees and Basin Hopping Yields an Improved Exploration of Energy Landscapes**
Reference: [22]

In a nutshell: Energy landscapes of biomolecular systems code their emergent thermodynamic and kinetic properties, so that their exploration is a question of paramount importance. This task requires in particular finding (metastable) states and their occupancy probabilities. Landscape exploration methods can be ascribed to two categories: continuous methods related to molecular dynamics, and discrete methods related to Monte Carlo sampling.

Assessment: In this work, we present a discrete sampling method combining features of robotics inspired methods (rapidly expanding random trees), and of biophysics inspired methods (basin hopping). Our hybrid algorithm outperforms contenders significantly. It is possibly one of the most efficient sampling method for energy landscapes known to date, but making such a statement will require testing thoroughly on a variety of systems. The method may strike a major impact if we manage to qualify the conformational ensembles generated from a thermodynamic standpoint.

► **Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison**
Reference: [16]

In a nutshell: A paper presenting novel methods to analyze conformational ensembles and sampled energy landscapes, using techniques from optimal transportation theory and computational topology.

Assessment: The method proposed significantly enriches those classically used in biophysics, and triggered a collaboration with David Wales (Cambridge), one of the leading scientists on energy landscapes.

4.1.3. The Structural Bioinformatics Library

We released the Structural Bioinformatics Library, a library whose main features are detailed below.

5. New Software and Platforms

5.1. The Structural Bioinformatics Library

5.1.1. Web site:

http://sbl.inria.fr

5.1.2. The SBL : Overview

The SBL is a generic C++/python library providing algorithms and applications to solve complex problems in computational structural biology (CSB).

For Biologists, the key advantages are:

- comprehensive in silico environment providing applications,
- answering complex bio-physical problems,
- 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.

5.1.3. 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.

5.1.4. 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.

5.1.5. 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. In a nutshell, concepts are a type system for types, and models are specific classes following this system.

5.1.6. SBL-CORE: the SBL for Low-level Developers and Contributors

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 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.

5.1.7. 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/).

5.1.8. 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. High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions

Participants: Frédéric Cazals, Simon Marillet.

In collaboration with P. Boudinot, Unité de recherche en virologie et immunologie moléculaires, INRA Jouy-en-Josas.

Predicting protein binding affinities from structural data has remained elusive, a difficulty owing to the variety of protein binding modes. Using the structure-affinity-benchmark (SAB, 144 cases with bound/unbound crystal structures and experimental affinity measurements), prediction has been undertaken either by fitting a model using a handfull of pre-defined variables, or by training a complex model from a large pool of parameters (typically hundreds). The former route unnecessarily restricts the model space, while the latter is prone to overfitting.

We design models in a third tier [20], using twelve variables describing enthalpic and entropic variations upon binding, and a model selection procedure identifying the best sparse model built from a subset of these variables. Using these models, we report three main results. First, we present models yielding a marked improvement of affinity predictions. For the whole dataset, we present a model predicting K_d within one and two orders of magnitude for 48% and 79% of cases, respectively. These statistics jump to 62% and 89% respectively, for the subset of the SAB consisting of high resolution structures. Second, we show that these performances owe to a new parameter encoding interface morphology and packing properties of interface atoms. Third, we argue that interface flexibility and prediction hardness do not correlate, and that for flexible cases, a performance matching that of the whole SAB can be achieved. Overall, our work suggests that the affinity prediction problem could be partly solved using databases of high resolution complexes whose affinity is known.

6.1.2. *Dissecting Interfaces of Antibody - Antigen Complexes: from Ligand Specific Features to Binding Affinity Predictions*

Participants: Frédéric Cazals, Simon Marillet.

In collaboration with: P. Boudinot, Unité de recherche en virologie et immunologie moléculaires, INRA Jouy-en-Josas; M-P. Lefranc, Univ. of Montpellier 2.

B lymphocytes recognize the antigen through their membrane immunoglobulins (IG), that can also be secreted. The diversity of IG-Ag complexes challenges our understanding in terms of binding affinity and interaction specificity.

In this work [21], we dissect the interfaces of IG-Ag complexes from high resolution crystal structures. We show that global interface statistics distinguish ligand types and that interfacial side chains play a key role in the interaction. Our analysis of the relative positions of CDR identifies a remarkably conserved pattern involving seven seams between CDR, with specific variations depending on the ligand type. Finally, we show that structural features of the interface and of the partners yield binding affinity estimates of unprecedented accuracy (median absolute error of 1.02 kcal/mol).

Our findings will be of broad interest, as understanding Ag recognition at the atomic level will help guiding design of better IG targeting Ag for therapeutic or other uses.

6.2. Modeling Macro-molecular Assemblies

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

6.2.1. *Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems*

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 context, we present two novel algorithms, `MILP-W` and `MILP-W`$_B$ [14]. 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 in the `SBL` , and should prove of ubiquitous interest whenever connectivity inference from oligomers is faced.

6.3. Modeling the Flexibility of Macro-molecules

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

6.3.1. Hybridizing Rapidly Growing Random Trees and Basin Hopping Yields an Improved Exploration of Energy Landscapes

Participants: Frédéric Cazals, Tom Dreyfus, Christine Roth.

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

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 [22], 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. Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces

Participants: Frédéric Cazals, Alix Lhéritier.

Comparing two sets of multivariate samples is a central problem in data analysis. From a statistical standpoint, the simplest way to perform such a comparison is to resort to a non-parametric two-sample test (TST), which checks whether the two sets can be seen as i.i.d. samples of an identical unknown distribution (the null hypothesis). If the null is rejected, one wishes to identify regions accounting for this difference. In this paper [17], we presents a two-stage method providing *feedback* on this difference, based upon a combination of statistical learning (regression) and computational topology methods.

Consider two populations, each given as a point cloud in $\mathbb{R}^d$. In the first step, we assign a label to each set and we compute, for each sample point, a discrepancy measure based on comparing an estimate of the conditional probability distribution of the label given a position versus the global unconditional label distribution. In the second step, we study the height function defined at each point by the aforementioned estimated discrepancy. Topological persistence is used to identify persistent local minima of this height function, their *basins* defining regions of points with high discrepancy and in spatial proximity.

Experiments are reported both on synthetic and real data (satellite images and handwritten digit images), ranging in dimension from $d = 2$ to $d = 784$, illustrating the ability of our method to localize discrepancies.

On a general perspective, the ability to provide feedback downstream TST may prove of ubiquitous interest in exploratory statistics and data science.

6.4.2. A Sequential Non-parametric Two-Sample Test

Participants: Frédéric Cazals, Alix Lhéritier.

Given samples from two distributions, a nonparametric two-sample test aims at determining whether the two distributions are equal or not, based on a test statistic. This statistic may be computed on the whole dataset, or may be computed on a subset of the dataset by a function trained on its complement. We propose a third tier [19], consisting of functions exploiting a sequential framework to learn the differences while incrementally processing the data. Sequential processing naturally allows optional stopping, which makes our test the first truly sequential nonparametric two-sample test.

We show that any sequential predictor can be turned into a sequential two-sample test for which a valid p-value can be computed, yielding controlled type I error. We also show that pointwise universal predictors yield consistent tests, which can be built with a nonparametric regressor based on k-nearest neighbors in particular. We also show that mixtures and switch distributions can be used to increase power, while keeping consistency.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. 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.

7.2. International Research Visitors

7.2.1. Visits of International Scientists

7.2.1.1. Internships

- N. Gayraud, 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 *Modeling cryo-electron microscopy maps*. Nathalie is now following-up as a PhD student in the Athena project team.

- S. Lundy (Supélec, Gif-sur-Yvette), completed a 3 months internship under the joint supervision of Dorian Mazauric and Jean-Daniel Boissonnat (Geometrica, Inria Sophia Antipolis Méditerranée) on the topic *Representation of simplicial complexes by directed graphs*.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events Organisation

8.1.1.1. General Chair, Scientific Chair

Together with J. Cortés (LAAS/CNRS, Toulouse), and C. Robert (IBPC/CNRS, Paris), we launched and have been organizing the Winter Schools series *Algorithms in Structural Bio-informatics*. These schools are meant to train PhD students and post-docs on advanced algorithmic techniques in structural biology. The 2015 Edition, who took place at the CNRS center in Cargese, focused on *Sampling bio-molecular systems*, see http://algosb.galaxy.ibpc.fr/.

8.1.2. Scientific Events Selection

8.1.2.1. Member of the Conference Program Committees

F. Cazals was member of the following program committees:

- Symposium On Geometry Processing

8.1.3. Invited Talks

– F. Cazals gave the following invited talks:

- *Energy Landscapes: Sampling, Analysis, and Comparison*, Max-Planck Institute for Solid State Research, Stuttgart, Germany. November 2015.
- *Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces* GUDHI workshop on topological data analysis, Porquerolles, France. October 2015.
- *Exploring and modeling energy landscapes*, University Chemical Laboratories, Cambridge University, UK. February 2015.

– D. Mazauric gave the following invited talks:

- *Mass Transportation Problems with Connectivity Constraints and Energy Landscape Comparison*, Laboratoire d'Informatique Fondamentale de Marseille. March 2015.
- *Representation of simplicial complexes*, Winter School on Algorithmic Geometry of Triangulations, Inria Sophia Antipolis - Méditerranée. January 2015.

8.1.4. Leadership within the Scientific Community

– F. Cazals:

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

8.2. Teaching - Supervision - Juries

8.2.1. 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.

8.2.2. Supervision

(PhD thesis, defended, November 2015) A. Lhéritier, *Nonparametric methods for learning and detecting multivariate statistical dissimilarity*, University of Nice Sophia Antipolis. Advisor: F. Cazals.

(PhD thesis, defended, May 2015) D. Agarwal, *Topics in mass spectrometry based structure determination*, University of Nice Sophia Antipolis. Advisor: F. Cazals.

(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) 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.

(PhD thesis, ongoing) A. Chevalier, *Sampling biomolecular systems*, University of Nice Sophia Antipolis.

8.2.3. Juries

– F. Cazals:

- Mathilde Le Boudic-Jamin, University of Rennes 1, December 2015. Rapporteur on the PhD thesis *Similarités et divergences, globales et locales, entre structures protéiques*. Advisor: R. Andonov.
- Nathan Desdouits, University Pierre et Marie Curie / Institut Pasteur Paris, May 2015. Rapporteur on the PhD thesis *Concepts et méthodes d'analyse numérique de la dynamique des cavités au sein des protéines applications à l'élaboration de stratégies novatrices d'inhibition*. Advisors: Michael Nilges and Arnaud Blondel.
- Petr Popov, University of Grenoble, January 2015. Committee member. *New methods for the prediction of protein - protein interactions at the structural level*. Advisor: Sergei Grudinin.

8.3. Popularization

Dorian Mazauric. Dorian is a member of the group of Médiation et Animation des MAthématiques, des Sciences et Techniques Informatiques et des Communications (MASTIC), Inria Sophia Antipolis - Méditerranée. He participated to the following events:

- 12-16/10/2015: Fête de la Science at collège Yves Montand, Vinon-sur-Verdon. *Graphes et algorithmes pour tous*. Organized by Institut Esope 21.
- 13/10/2015: Conference at collège Yves Montand, Vinon-sur-Verdon (avec Frédéric Havet, COATI project-team). *Présentation du métier de chercheur*. Organized by Institut Esope 21.
- 10-11/10/2015: Village des sciences et de l'innovation au Palais des Congrès d'Antibes Juan-les-Pins. Fête de la Science 2015. *Graphes et algorithmes pour tous*.

- 17/06/2015: Stage MathC2+ à Inria Sophia Antipolis - Méditerranée. *Théorie des graphes et algorithmique*.
- 18/03/2015: Conference at lycée Henri Matisse de Vence. *La théorie des graphes et ses applications dans les réseaux*. Dans le cadre du dispositif régional "Science Culture".
- 17/12/2015: Presentation at école élémentaire Sartoux, Sophia Antipolis (classe de CM2). *La magie des graphes et du binaire, algorithmes et jeux*. This presentation is part of the cycle ASTEP (Accompagnement en Sciences et Technologies à l'École Primaire).
- 14/12/2015: Presentation at école élémentaire de Montaleigne, Saint-Laurent-du-Var (five classes). *La magie des graphes et du binaire, algorithmes et jeux*. This presentation is part of the cycle ASTEP (Accompagnement en Sciences et Technologies à l'École Primaire).

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] 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

[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. *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

[6] 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

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

[8] 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

[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

Doctoral Dissertations and Habilitation Theses

[12] D. AGARWAL. *Topics in mass spectrometry based structure determination*, Université Nice Sophia Antipolis, May 2015, https://tel.archives-ouvertes.fr/tel-01176554

[13] A. LHÉRITIER. *Nonparametric Methods for Learning and Detecting Multivariate Statistical Dissimilarity*, Université Nice Sophia Antipolis, November 2015, https://hal.inria.fr/tel-01245946

Articles in International Peer-Reviewed Journals

[14] 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", April 2015, 27 p. [*DOI :* 10.1074/MCP.M114.047779], https://hal.inria.fr/hal-01245401

[15] 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

[16] F. CAZALS, T. DREYFUS, D. MAZAURIC, A. ROTH, C. ROBERT. *Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison*, in "Journal of Computational Chemistry", May 2015, vol. 36, n^{o} 6, 18 p. , https://hal.archives-ouvertes.fr/hal-01245395

International Conferences with Proceedings

[17] F. CAZALS, A. LHÉRITIER. *Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces*, in "IEEE/ACM International Conference on Data Science and Advanced Analytics", Paris, France, P. GALLINARI, J. KWOK, G. PASI, O. ZAIANE (editors), October 2015, 29 p. , https://hal.inria.fr/hal-01245408

Research Reports

[18] F. CAZALS, A. LHÉRITIER. *Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces*, Inria, March 2015, n^{o} RR-8734, 29 p. , https://hal.inria.fr/hal-01159235

[19] A. LHÉRITIER, F. CAZALS. *A Sequential Nonparametric Two-Sample Test*, Inria, March 2015, n^{o} RR-8704, 18 p. , https://hal.inria.fr/hal-01135608

[20] S. MARILLET, P. BOUDINOT, F. CAZALS. *High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions*, Inria, March 2015, n^{o} RR-8733, https://hal.inria.fr/hal-01159641

[21] S. Marillet, M.-P. Lefranc, P. Boudinot, F. Cazals. *Dissecting Interfaces of Antibody -Antigen Complexes: from Ligand Specific Features to Binding Affinity Predictions*, Inria Sophia Antipolis, September 2015, n^{o} RR-8770, 61 p. , https://hal.inria.fr/hal-01191462

[22] A. Roth, T. Dreyfus, C. H. Robert, F. Cazals. *Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes*, Inria, September 2015, n^{o} RR-8768, 29 p. , https://hal.inria.fr/hal-01191028

References in notes

[23] 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

[24] 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

[25] 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

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

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

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

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

[30] 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

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

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

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

[40] 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

[41] 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

[42] 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

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

[44] 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

[45] 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

[46] 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

[47] 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

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