Activity Report 2017

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:

A2.5. - Software engineering
A3.3.2. - Data mining
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A6.1.4. - Multiscale modeling
A6.2.4. - Statistical methods
A6.2.8. - Computational geometry and meshes
A8.1. - Discrete mathematics, combinatorics
A8.3. - Geometry, Topology
A8.7. - Graph theory
A9.2. - Machine learning

Other Research Topics and Application Domains:

B1.1.1. - Structural biology
B1.1.7. - Immunology
B1.1.9. - Bioinformatics

1. Personnel

Research Scientists

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

Post-Doctoral Fellow

Rémi Watrigant [Inria, until Sep 2017]

PhD Students

Denys Bulavka [Inria, from Oct 2017]
Augustin Chevallier [Université Côte d'Azur]
Méliné Simsir [Université Côte d'Azur, from Dec 2017]
Romain Tetley [Université Côte d'Azur]

Intern

Louis Becquey [Inria, from May 2017 until Aug 2017]

Administrative Assistant

Florence Barbara [Inria]

External Collaborators

Neva Durand [Baylor college of medicine, Visiting scientist, until Jul 2017]
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.

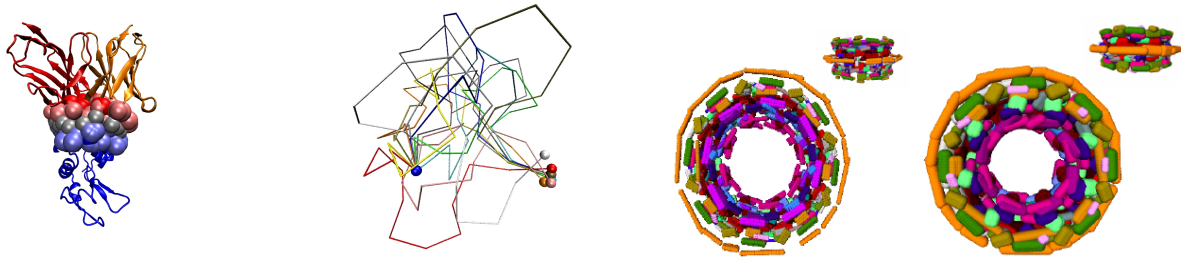
Ideally, bio-physical models of macro-molecules should resort to quantum mechanics. While this is possible for small systems, say up to 50 atoms, large systems are investigated within the framework of the Born-Oppenheimer approximation which stipulates the nuclei and the electron cloud can be decoupled. 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. This model is instrumental in mining correlations between structural and biological as well as biophysical properties of protein complexes [12]. (b) A diverse set of conformations of a backbone loop, selected thanks to a geometric optimization algorithm [7]. 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 [12]. To elaborate at the atomic level, one seeks a structural alphabet encoding the spatial structure of proteins. At the side-chain and backbone level, an example of such alphabet is that of [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. 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. SBL

Structural Bioinformatics Library
KEYWORDS: Structural Biology - Biophysics - Software architecture
FUNCTIONAL DESCRIPTION: The SBL is a generic C++/python cross-platform software library targeting complex problems in structural bioinformatics. Its tenet is based on a modular design offering a rich and versatile framework allowing the development of novel applications requiring well specified complex operations, without compromising robustness and performances.

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

RELEASE FUNCTIONAL DESCRIPTION: In 2017, major efforts targeted two points. First, the simplification of installation procedures. Second, the development of packages revolving on molecular flexibility at large: representations in internal and Cartesian coordinates, generic representation of molecular mechanics force fields (and computation of gradients), exploration algorithms for conformational spaces.

- Contact: Frédéric Cazals
- Publication: The Structural Bioinformatics Library: modeling in biomolecular science and beyond
- URL: https://sbl.inria.fr/

5. New Results

5.1. Modeling interfaces and contacts

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

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

Participants: F. Cazals, S. Marillet.

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

Antibody-antigen complexes challenge our understanding, as analyses to date failed to unveil the key determinants of binding affinity and interaction specificity. In this work [17], we partially fill this gap based on novel quantitative analyses using two standardized databases, the IMGT/3Dstructure-DB and the structure affinity benchmark.

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

5.1.2. Anti-interleukin-6 signalling therapy rebalances the disrupted cytokine production of B cells from patients with active rheumatoid arthritis

Participants: F. Cazals, A. Lhéritier.

In collaboration with S. Fleischer (1. Charité University Medicine Berlin, Berlin, Germany), S. Ries (2. Deutsches Rheuma-Forschungszentrum Berlin, Berlin, Germany), P. Shen (2.), G.R. Burmester (1.), T. Dörner (1.), S. Fillatreau (2., Institut Necker-Enfants Malades, Université Paris Descartes, IHP Hôpital Necker Enfants Malades).

Rheumatoid arthritis (RA) is associated with abnormal B cell-functions implicating antibody-dependent and -independent mechanisms. B cells have emerged as important cytokine-producing cells, and cytokines are well-known drivers of RA pathogenesis. To identify novel cytokine-mediated B-cell functions in RA, in this work [16], we comprehensively analysed the capacity of B cells from RA patients with an inadequate response to disease modifying anti-rheumatic drugs to produce cytokines in comparison with healthy donors (HD). RA B cells displayed a constitutively higher production of the pathogenic factors interleukin (IL)-8 and Gro-α, while their production of several cytokines upon activation via the B cell receptor for antigen (BCR) was broadly suppressed, including a loss of the expression of the protective factor TRAIL, compared to HD B cells. These defects were partly erased after treatment with the IL-6-signalling inhibitor tocilizumab, indicating that abnormal IL-6 signalling contributed to these abnormalities. Noteworthy, the clinical response of individual patients to tocilizumab therapy could be predicted using the amounts of MIP-1β and β-NGF produced by these patients' B cells before treatment. Taken together, our study highlights hitherto unknown abnormal B-cell functions in RA patients, which are related to the unbalanced cytokine network, and are potentially relevant for RA pathogenesis and treatment.

5.2. Modeling macro-molecular assemblies

Keywords: macro-molecular assembly, reconstruction by data integration, proteomics, mass spectrometry, modeling with uncertainties, connectivity inference.

5.2.1. Complexity dichotomies for the minimum F-overlay problem

Participants: D. Mazauric, R. Watrigant.

In collaboration with N. Cohen (LRI, UMR de l'Université Paris-Sud et du CNRS), F. Havet (Université Côte d'Azur, I3S, UMR de l'Université Nice Sophia et du CNRS), I. Sau (LIRMM, UMR de l'Université Montpellier et du CNRS, and Universidade Federal do Ceará, Brazil).

The *connectivity inference* problem for native mass spectrometry aims at finding the most plausible pairwise contacts between the individual subunits of a macro-molecular assembly, given the composition of overlapping oligomers. The associated combinatorial optimization problem consists in determining a minimal-cardinality set of contact (edges) such that all the subunits of each oligomer must be "connected" (each oligomer must induce a connected graph). We studied in [18] the general inference problem that consists of considering more general properties on oligomers. For this new problem, we are given a list of possible topologies (graphs) for each oligomer and we aim at minimizing the total number of contacts between subunits. In terms of graphs, we are given a family of subgraphs that can match the structure of the oligomers. These new constraints reflect biophysical properties: a subunit has a limited number of neighbors (bounded maximum degree of the subgraphs), selected contacts are already known (a given subgraph contained in the complex), etc. We prove that the problem is NP-complete (no polynomial time algorithm, unless P = NP) for almost all cases.

5.3. Modeling the flexibility of macro-molecules

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

No new result on this topic in 2017.

5.4. 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 discussed below.

5.4.1. Extracting the groupwise core structural connectivity network: bridging statistical and graph-theoretical approaches

Participant: D. Mazauric.

In collaboration with N. Lascano (Universidad de Buenos Aires, Argentina, Université Côte d'Azur, and Inria Sophia Antipolis - Méditerranée, EPI ATHENA), G. Gallardo (2. Université Côte d'Azur and Inria Sophia Antipolis - Méditerranée, EPI ATHENA), D. Wassermann (2).

Finding the common structural brain connectivity network for a given population is an open problem, crucial for current neuro-science. Recent evidence suggests there is a tightly connected network shared between humans. Obtaining this network will, among many advantages, allow us to focus cognitive and clinical analyses on common connections, thus increasing their statistical power. In turn, knowledge about the common network will facilitate novel analyses to understand the structure-function relationship in the brain. In [19], we present a new algorithm for computing the core structural connectivity network of a subject sample combining graph theory and statistics. Our algorithm works in accordance with novel evidence on brain topology. We analyze the problem theoretically and prove its complexity. Using 309 subjects, we show its advantages when used as a feature selection for connectivity analysis on populations, outperforming the current approaches.

5.4.2. Maximum flow under proportional delay constraint

Participant: D. Mazauric.

In collaboration with P. Bonami (LIF, UMR d'Aix-Marseille Université et du CNRS, and IBM ILOG CPLEX, Madrid), Y. Vaxès (LIF, UMR d'Aix-Marseille Université et du CNRS).

Network operators must satisfy some Quality of Service requirements for their clients. One of the most important parameters in telecommunication networks is the end-to-end delay of a unit of flow between a source node and a destination node. Given a network and a set of source destination pairs (connections), we consider in [14] the problem of maximizing the sum of the flow under proportional delay constraints. In this paper, the delay for crossing a link is proportional to the total flow crossing this link. If a connection supports non-zero flow, then the sum of the delays along any path corresponding to that connection must be lower than a given bound. The constraints of delay are on-off constraints because if a connection carries zero flow, then there is no constraint for that connection. The difficulty of the problem comes from the choice of the connections supporting non-zero flow. We first prove a general approximation ratio using linear programming for a variant of the problem. We then prove a linear time 2-approximation algorithm when the network is a path. We finally show a Polynomial Time Approximation Scheme when the graph of intersections of the paths has bounded treewidth.

5.4.3. Comparing two clusterings using matchings between clusters of clusters

Participants: F. Cazals, D. Mazauric, R. Tetley, R. Watrigant.

Clustering is a fundamental problem in data science, yet, the variety of clustering methods and their sensitivity to parameters make clustering hard. To analyze the stability of a given clustering algorithm while varying its parameters, and to compare clusters yielded by different algorithms, several comparison schemes based on matchings, information theory and various indices (Rand, Jaccard) have been developed. In this work [20], we go beyond these by providing a novel class of methods computing meta-clusters within each clustering– a meta-cluster is a group of clusters, together with a matching between these. Let the intersection graph of two clusterings be the edge-weighted bipartite graph in which the nodes represent the clusters, the edges represent the non empty intersection between two clusters, and the weight of an edge is the number of common items. We introduce the so-called D-family-matching problem on intersection graphs, with D the upper-bound on the diameter of the graph induced by the clusters of any meta-cluster. First we prove NP-completeness results and unbounded approximation ratio of simple strategies. Second, we design exact polynomial time dynamic programming algorithms for some classes of graphs (in particular trees). Then, we prove spanning-tree based efficient algorithms for general graphs. Our experiments illustrate the role of D as a scale parameter providing information on the relationship between clusters within a clustering and in-between two clusterings. They also show the advantages of our built-in mapping over classical cluster comparison measures such as the variation of information (VI).

5.4.4. The SBL

Participants: F. Cazals, T. Dreyfus.

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

The SBL involves four software components (1–4 thereafter) [15]. For end-users, the SBL provides ready to use, state-of-the-art (1) applications to handle molecular models defined by unions of balls, to deal with molecular flexibility, to model macro-molecular assemblies. These applications can also be combined to tackle integrated analysis problems. For developers, the SBL provides a broad C++ toolbox with modular design, involving core (2) algorithms, (3) biophysical models and (4) modules, the latter being especially suited to develop novel applications. The SBL comes with a thorough documentation consisting of user and reference manuals, and a bugzilla platform to handle community feedback.

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

See also the section **New Software and Platforms**.

6. Bilateral Contracts and Grants with Industry

6.1. Bilateral Contracts with Industry

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

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

6.1.1. Context

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

6.1.2. Specific goals

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

7. Dissemination

7.1. Promoting Scientific Activities

7.1.1. 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 2017 Edition, which took place at the CNRS center in Cargese, focused on *Protein design*, see https://algosb2017.sciencesconf.org/. It attracted 40 participants from worldwide institutions.

7.1.2. Scientific Events Selection

7.1.2.1. Member of Conference Program Committees

– Frédéric Cazals was member of the following program committees:

- Symposium On Geometry Processing
- Symposium on Solid and Physical Modeling
- Intelligent Systems for Molecular Biology (ISMB), PC member of Protein Interactions & Molecular Networks
- ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics
- JOBIM

7.1.3. Journal

7.1.3.1. Reviewer - Reviewing Activities

- Journal of mathematical biology
- Bioinformatics
- Journal of computational chemistry

7.1.4. Invited Talks

– Frédéric Cazals gave the following invited talks:

- *Beyond two-sample-tests: localizing data discrepancies in high-dimensional spaces*, NIPS workshop of Topological Data Analysis, Los Angeles, December 2017.
- *Modeling in structural bioinformatics: the tryptic structure - dynamics - function*, GDR Bioinformatique moléculaire, Paris, November 2017.
- *Modèles géométriques pour la prédiction des interactions macro-moléculaires*, seminar for the course *Géométrie algorithmique Données, Modèles, Programmes*, by Jean-Daniel Boissonnat, Chaire d'informatique et sciences numériques, Collège de France, March 2017.

7.1.5. Leadership within the Scientific Community

– Frédéric Cazals:

- 2010-.... Member of the steering committee of the *GDR Bioinformatique Moleculaire*, for the *Structure and macro-molecular interactions* theme.
- 2017-.... Co-chair, with Yann Ponty, of the working group / groupe de travail *(GT MASIM - Méthodes Algorithmiques pour les Structures et Interactions Macromoléculaires*, within the *GDR de BIoinformatique Moléculaire* (GDR BIM, http://www.gdr-bim.cnrs.fr/).

7.1.6. Research Administration

– Frédéric Cazals:

- 2017-.... President of the *Comité de suivi doctoral* (CSD), Inria Sophia Antipolis - Méditerranée. The CSD supervises all aspects of PhD student's life within Inria sophia antipolis.

– Dorian Mazauric:

- 2016-2019. Member of the *Comité de Centre*, Inria Sophia Antipolis - Méditerranée.

7.2. Teaching - Supervision - Juries

7.2.1. Teaching

Master: Frédéric Cazals (Inria ABS) and Frédéric Chazal (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)

7.2.2. Supervision

PhD thesis, ongoing, 3rd year. Romain Tetley, *Structural alignments: beyond the rigid case*. Université Côte d'Azur. Under the supervision of Frédéric Cazals.

PhD thesis, ongoing, 3rd year. Augustin Chevallier, *Sampling biomolecular systems*. Université Côte d'Azur. Under the supervision of Frédéric Cazals.

PhD thesis, ongoing, 1st year. Denys Bulavka, *Modeling macro-molecular motions*. Université Côte d'Azur. Under the supervision of Frédéric Cazals.

PhD thesis, ongoing, 1st year. Méliné Simsir, *Modeling drug efflux by Patched*. Université Côte d'Azur. Thesis co-supervised by Frédéric Cazals and Isabelle Mus-Veteau, IPMC/CNRS.

Postdoctoral research of Rémi Watrigant, 2016 - 2017. Projet de Recherche Exploratoire (Inria). *Improving inference algorithms for macromolecular structure determination*. Under the supervision of Dorian Mazauric and Frédéric Havet (Inria COATI project-team).

7.2.3. Juries

– Frédéric Cazals:

- Clément Viricel, University of Toulouse, December 2017. Rapporteur on the PhD thesis *Contributions au développement d'outils computationnels de design de protéines : méthodes et algorithmes de comptage avec garantie*. Advisors: T. Schiex and S. Barbe.

7.3. Popularization

This section describes the activities of Dorian Mazauric (member of the popularization committee of Inria Sophia Antipolis - Méditerranée).

– **Founding**.

- Coordinator of the project GALEJADE (*Graphes et ALgorithmes : Ensembles de Jeux À Destination des Écoliers*) founded by Fondation Blaise Pascal (2017 - 2018). We aim at developing an educational kit for primary schools in order to play with the notions of graphs and algorithms. We also propose conferences for the general public.

– **Resources**.

- 13 posters - *Transmission de pensée - La magie du binaire*: [22].
- 2 posters - *Tour de cartes - La magie des graphes et du binaire*: [21].

– **Activities**.

- Training with Laurent Giauffret of 30 teachers (cycle 1) of Académie de Nice (*Jeux avec les graphes et les algorithmes*).
- Stage MathC2+: half day activities for 40 high school students (Boruvka algorithm for the minimum spanning tree problem played on "real" graphs constructed with plastic hoops and slats).
- Fête de la Science 2017
  - Village des sciences et de l'innovation au Palais des Congrès d'Antibes Juan-les-Pins.
  - Conferences (classe préparatoire, 2 classes de seconde, une classe de sixième, 2 classes de CM1).
- 4 conferences in high schools (Aix-en-Provence, Antibes, Grasse, Miramas), dispositif Science Culture PACA.
- One conference in a secondary school (French *college*) (Cagnes-sur-Mer).

8. Bibliography

Major publications by the team in recent years

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

[2] D. AGARWAL, J. ARAUJO, C. CAILLOUET, F. CAZALS, D. COUDERT, S. PÉRENNES. *Connectivity Inference in Mass Spectrometry based Structure Determination*, in "European Symposium on Algorithms (Springer LNCS 8125)", Sophia Antipolis, France, H. BODLAENDER, G. ITALIANO (editors), Springer, 2013, pp. 289–300, http://hal.inria.fr/hal-00849873

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

[4] J. CARR, D. MAZAURIC, F. CAZALS, D. J. WALES. *Energy landscapes and persistent minima*, in "The Journal of Chemical Physics", 2016, vol. 144, n° 5, 4 p. [*DOI :* 10.1063/1.4941052], https://www.repository.cam.ac.uk/handle/1810/253412

[5] 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

[6] F. CAZALS, T. DREYFUS, D. MAZAURIC, A. ROTH, C. ROBERT. *Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison*, in "J. of Computational Chemistry", 2015, vol. 36, n° 16, pp. 1213–1231 [*DOI :* 10.1002/JCC.23913], https://hal.archives-ouvertes.fr/hal-01076317

[7] F. CAZALS, T. DREYFUS, S. SACHDEVA, N. SHAH. *Greedy Geometric Algorithms for Collections of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining*, in "Computer Graphics Forum", 2014, vol. 33, n° 6, pp. 1–17 [*DOI :* 10.1111/CGF.12270], http://hal.inria.fr/hal-00777892

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

[9] T. DREYFUS, V. DOYE, F. CAZALS. *Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models*, in "Proteins: structure, function, and bioinformatics", 2012, vol. 80, n° 9, pp. 2125–2136

[10] T. DREYFUS, V. DOYE, F. CAZALS. *Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Sub-complexes*, in "Proteins: structure, function, and bioinformatics", 2013, vol. 81, n° 11, pp. 2034–2044 [*DOI :* 10.1002/PROT.24313], http://hal.inria.fr/hal-00849795

[11] N. MALOD-DOGNIN, A. BANSAL, F. CAZALS. *Characterizing the Morphology of Protein Binding Patches*, in "Proteins: structure, function, and bioinformatics", 2012, vol. 80, n° 12, pp. 2652–2665

[12] S. MARILLET, P. BOUDINOT, F. CAZALS. *High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions*, in "Proteins: structure, function, and bioinformatics", 2015, vol. 1, n° 84, pp. 9–20 [*DOI :* 10.1002/PROT.24946], https://hal.inria.fr/hal-01159641

[13] A. ROTH, T. DREYFUS, C. ROBERT, F. CAZALS. *Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes*, in "J. Comp. Chem.", 2016, vol. 37, n° 8, pp. 739–752 [*DOI :* 10.1002/JCC.24256], https://hal.inria.fr/hal-01191028

Publications of the year

Articles in International Peer-Reviewed Journals

[14] P. BONAMI, D. MAZAURIC, Y. VAXÈS. *Maximum flow under proportional delay constraint*, in "Theoretical Computer Science", 2017, vol. 689, pp. 58-66 [*DOI :* 10.1016/J.TCS.2017.05.034], https://hal.inria.fr/hal-01571232

[15] F. CAZALS, T. DREYFUS. *The Structural Bioinformatics Library: modeling in biomolecular science and beyond*, in "Bioinformatics", April 2017, vol. 33, n° 8 [*DOI :* 10.1093/BIOINFORMATICS/BTW752], https://hal.inria.fr/hal-01570848

[16] S. FLEISCHER, S. RIES, P. SHEN, A. LHÉRITIER, F. CAZALS, G. R. BURMESTER, T. DÖRNER, S. FILLATREAU. *Anti-interleukin-6 signalling therapy rebalances the disrupted cytokine production of B cells from patients with active rheumatoid arthritis*, in "European Journal of Immunology", September 2017 [*DOI :* 10.1002/EJI.201747191], https://hal.inria.fr/hal-01671956

[17] S. MARILLET, M.-P. LEFRANC, P. BOUDINOT, F. CAZALS. *Novel Structural Parameters of Ig–Ag Complexes Yield a Quantitative Description of Interaction Specificity and Binding Affinity*, in "Frontiers in Immunology", February 2017, vol. 8 [*DOI :* 10.3389/FIMMU.2017.00034], https://hal.inria.fr/hal-01570846

International Conferences with Proceedings

[18] N. COHEN, F. HAVET, D. MAZAURIC, I. SAU, R. WATRIGANT. *Complexity Dichotomies for the Minimum F -Overlay Problem*, in "IWOCA 2017 - 28th International Workshop on Combinatorial Algorithms", Newcastle, Australia, July 2017, 12 p. , https://hal.inria.fr/hal-01571229

[19] N. LASCANO, G. GALLARDO, R. DERICHE, D. MAZAURIC, D. WASSERMANN. *Extracting the Groupwise Core Structural Connectivity Network: Bridging Statistical and Graph-Theoretical Approaches*, in "Information Processing in Medical Imaging", Boone, United States, 2017, https://arxiv.org/abs/1701.01311 , https://hal.inria.fr/hal-01426870

Research Reports

[20] F. CAZALS, D. MAZAURIC, R. TETLEY, R. WATRIGANT. *Comparing two clusterings using matchings between clusters of clusters*, Inria Sophia Antipolis - Méditerranée ; Universite Cote d'Azur, April 2017, n° RR-9063, https://hal.inria.fr/hal-01514872

Scientific Popularization

[21] D. MAZAURIC. *Tour de cartes - La magie des graphes et du binaire*, 2017, 2 p. , Posters expliquant un tour de cartes qui utilise les graphes et le codage binaire, https://hal.inria.fr/hal-01671009

[22] D. MAZAURIC. *Transmission de pensée - La magie du binaire*, 2017, 13 p. , Posters expliquant le binaire avec un tour de magie, https://hal.inria.fr/hal-01670180

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° 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° 5, pp. 548–558

[29] R. DUNBRACK. *Rotamer libraries in the 21st century*, in "Curr Opin Struct Biol", 2002, vol. 12, n° 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° 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° 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° 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