Activity Report 2018

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
B1.1.5. - Immunology
B1.1.7. - Bioinformatics

1. Team, Visitors, External Collaborators

Research Scientists

Frédéric Cazals [Team leader, Inria, Senior Researcher, HDR]
Mehmet Serkan Apaydin [Inria, Starting Research Position, from Mar 2018]
Dorian Mazauric [Inria, Researcher]

PhD Students

Denys Bulavka [Inria]
Augustin Chevallier [Université Côte d'Azur, until Nov 2018]
Thi Viet Ha Nguyen [Inria, from Sep 2018]
Timothée O'Donnell [Inria, from Oct 2018]
Méliné Simsir [Université Côte d'Azur]
Romain Tetley [Université Côte d'Azur, until Dec 2018]

Interns

Maria Guramare [Harvard Universiy, from May 2018 until Jun 2018]
Thi Viet Ha Nguyen [Inria, from Mar 2018 until June 2018]
Timothée O'Donnell [Inria, from Mar 2018 until Aug 2018]
Xuchun Zhang [Inria, from Jul 2018 until Sep 2018]

Administrative Assistant

Florence Barbara [Inria]

Visiting Scientist

Marcin Pacholczyk [Silesian University of Technology, Poland, until Feb 2018]

External Collaborators

Charles Robert [CNRS, from Nov 2018, HDR]
Tom Dreyfus [RedHant Labs, from Apr 2018]

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 [52]. 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 [50], [38] and later Connolly [33], 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) [40], the number of distinct conformations of a poly-peptidic chain is exponential in the number of amino-acids. On the other hand, Nature folds proteins within time scales ranging from milliseconds to hours, while time-steps used in molecular dynamics simulations are of the order of the femto-second, so that biologically relevant time-scales are out reach for simulations. The fact that Nature avoids the exponential trap is known as Levinthal's paradox. The intrinsic difficulty of problems calls for models exploiting several classes of informations. For small systems, *ab initio* models can be built from first principles. But for more complex systems, *homology* or template-based models integrating a variable amount of knowledge acquired on similar systems are resorted to.

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

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

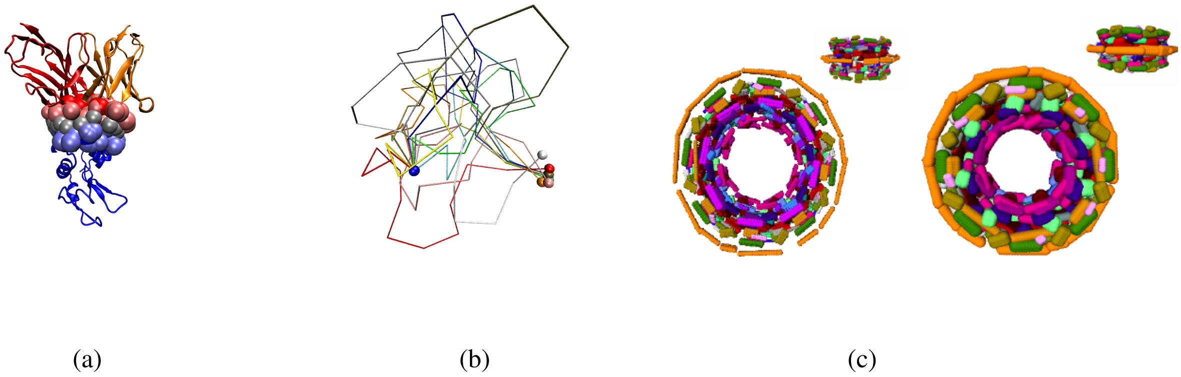

*Figure 1. **Geometric constructions in computational structural biology.** (a) An antibody-antigen complex, with interface atoms identified by our Voronoi based interface model. This model is instrumental in mining correlations between structural and biological as well as biophysical properties of protein complexes [12]. (b) A diverse set of conformations of a backbone loop, selected thanks to a geometric optimization algorithm [8]. 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 [52]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [55]. Current investigations follow two routes. From the experimental perspective [37], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [49]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [44].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change [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 [31], the usual way to design scoring functions is to mimic the so-called potentials of mean force. To put it briefly, one reverts Boltzmann's law, that is, denoting $p_i(r)$ the probability of two atoms –defining type i– to be located at distance r, the (free) energy assigned to the pair is computed as $E_i(r) = -kT \log p_i(r)$. Estimating from the PDB one function $p_i(r)$ for each type of pair of atoms, the energy of a complex is computed as the sum of the energies of the pairs located within a distance threshold [53], [39]. To compare the energy thus obtained to a reference state, one may compute $E = \sum_i p_i \log p_i / q_i$, with p_i the observed frequencies, and q_i the frequencies stemming from an a priori model [45]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

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

In the static setting, one seeks the minimalist geometric model providing a relevant bio-physical signal. A first step in doing so consists of identifying interface atoms, so as to relate the geometry and the bio-chemistry at the interface level [12]. To elaborate at the atomic level, one seeks a structural alphabet encoding the spatial structure of proteins. At the side-chain and backbone level, an example of such alphabet is that of [32]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [54], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond – a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [36].

In the dynamic setting, one wishes to understand whether selected (hot) residues exhibit specific dynamic properties, so as to serve as anchors in a binding process [48]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

[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.

3.3. Modeling macro-molecular assemblies

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

3.3.1. Reconstruction by Data Integration

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [30]. In a nutshell, the strategy is reminiscent from NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This strategy has been in particular used to propose plausible models of the Nuclear Pore Complex [29], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

3.3.2. Modeling with Uncertainties and Model Assessment

Reconstruction by data integration requires three ingredients. First, a parametrized model must be adopted, typically a collection of balls to model a protein with pseudo-atoms. Second, as in NMR, a functional measuring the agreement between a model and the data must be chosen. In [28], this functional is based upon *restraints*, namely penalties associated to the experimental data. Third, an optimization scheme must be selected. The design of restraints is notoriously challenging, due to the ambiguous nature and/or the noise level of the data. For example, Tandem Affinity Purification (TAP) gives access to a *pullout* i.e. a list of protein types which are known to interact with one tagged protein type, but no information on the number of complexes or on the stoichiometry of proteins types within a complex is provided. In cryo-EM, the envelope enclosing an assembly is often imprecisely defined, in particular in regions of low density. For immuno-EM labelling experiments, positional uncertainties arise from the microscope resolution.

These uncertainties coupled with the complexity of the functional being optimized, which in general is non convex, have two consequences. First, it is impossible to single out a unique reconstruction, and a set of plausible reconstructions must be considered. As an example, 1000 plausible models of the NPC were reported in [28]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

3.4. Modeling the flexibility of macro-molecules

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

Proteins in vivo vibrate at various frequencies: high frequencies correspond to small amplitude deformations of chemical bonds, while low frequencies characterize more global deformations. This flexibility contributes to the entropy thus the *free energy* of the system *protein - solvent*. From the experimental standpoint, NMR studies generate ensembles of conformations, called *conformers*, and so do molecular dynamics (MD) simulations. Of particular interest while investigating flexibility is the notion of correlated motion. Intuitively, when a protein is folded, all atomic movements must be correlated, a constraint which gets alleviated when the protein unfolds since the steric constraints get relaxed [3]. Understanding correlations is of special interest to predict the folding

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

pathway that leads a protein towards its native state. A similar discussion holds for the case of partners within a complex, for example in the third step of the *diffusion - conformer selection - induced fit* complex formation model.

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

At the side-chain level, the question of improving rotamer libraries is still of interest [34]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [51]. The standard methodology to analyze such a point cloud consists of resorting to normal modes. Recently, though, more elaborate methods resorting to more local analysis [47], to Morse theory [42] and to analysis of meta-stable states of time series [43] have been proposed.

3.5. Algorithmic foundations

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

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

3.5.1. Modeling Interfaces and Contacts

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

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

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

3.5.2. Modeling Macro-molecular Assemblies

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

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

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

3.5.3. Modeling the Flexibility of Macro-molecules

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

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

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 2018, major efforts targeted two points. First, the simplification of installation procedures – now possible with conda/python. 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. *Origin of Public Memory B Cell Clones in Fish After Antiviral Vaccination*

Participants: F. Cazals, S. Marillet.

In collaboration with S. Magadan, L. Jouneau, S. Marillet, P. Boudinot (INRA, Virologie et Immunologie Moléculaires, Université Paris-Saclay, Jouy-en-Josas, France); M. Puelma Touzel, T. Mora, A. Walczak (Laboratoire de Physique Théorique, CNRS, Sorbonne Université, and Ecole Normale Supérieure (PSL), Paris, France); W. Chaara, A. Six (Sorbonne Université, INSERM, UMR S 959, Immunology-Immunopathology - Immunotherapy (I3), Paris, France); E. Quillet (INRA, Génétique Animale et Biologie Intégrative, Université Paris-Saclay, Jouy-en-Josas, France); O. Sunyer (Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, United States); S. Fillatreau (INEM, INSERM U1151/CNRS UMR8253, Institut Necker-Enfants Malades, Faculté de Médecine Paris Descartes, Paris, France; Faculté de Médecine, Université Paris Descartes, Sorbonne Paris Cité, Paris, France; Assistance Publique des Hopitaux de Paris (AP-HP), Hopital Necker Enfants Malades, Paris, France).

Vaccination induces *publi*c antibody clonotypes common to all individuals of a species, that may mediate universal protection against pathogens. Only few studies tried to trace back the origin of these public B-cell clones. Here [16] we used Illumina sequencing and computational modeling to unveil the mechanisms shaping the structure of the fish memory antibody response against an attenuated Viral Hemorrhagic Septicemia rhabdovirus. After vaccination, a persistent memory response with a public VH5JH5 IgM component was composed of dominant antibodies shared among all individuals. The rearrangement model showed that these public junctions occurred with high probability indicating that they were already favored before vaccination due to the recombination process, as shown in mammals. In addition, these clonotypes were in the naive repertoire associated with larger similarity classes, composed of junctions differing only at one or two positions by amino acids with comparable properties. The model showed that this property was due to selective processes exerted between the recombination and the naive repertoire. Finally, our results showed that public clonotypes greatly expanded after vaccination displayed several VDJ junctions differing only by one or two amino acids with similar properties, highlighting a convergent response. The fish public memory antibody response to a virus is therefore shaped at three levels: by recombination biases, by selection acting on the formation of the pre-vaccination repertoire, and by convergent selection of functionally similar clonotypes during the response. We also show that naive repertoires of IgM and IgT have different structures and sharing between individuals, due to selection biases. In sum, our comparative approach identifies three conserved features of the antibody repertoire associated with public memory responses. These features were already present in the last common ancestors of fish and mammals, while other characteristics may represent species-specific solutions.

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 – Application for low resolution models of macro-molecular assemblies*

Participant: D. Mazauric.

In collaboration with N. Cohen (CNRS, Laboratoire de Recherche en Informatique) and F. Havet (CNRS, Inria/I3S project-team Coati) and I. Sau (CNRS, Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier) and R. Watrigant (University Lyon I, Laboratoire de l'Informatique du Parallélisme).

In this article [14], we analyze a generalization of the minimum connectivity inference problem (MCI). MCI models the computation of low-resolution structures of macro-molecular assemblies, based on data obtained by native mass spectrometry. The generalization studied in this article, allows us to consider more refined constraints for the characterization of low resolution models of large assemblies. We model this problem by using hypergraphs: for a (possibly infinite) fixed family of graphs F, we say that a graph G *overlays* F on a hypergraph H if $V(H)$ is equal to $V(G)$ and the subgraph of G induced by every hyperedge of H contains some member of F as a spanning subgraph. While it is easy to see that the complete graph on $|V(H)|$ overlays F on a hypergraph H whenever the problem admits a solution, the Minimum F-Overlay problem asks for such a graph with at most k edges, for some given $k \in \mathbb{N}$. This problem allows to generalize some natural problems which may arise in practice. For instance, if the family F contains all connected graphs, then Minimum F-Overlay corresponds to the MCI problem. Our main contribution is a strong dichotomy result regarding the polynomial vs. NP-complete status with respect to the considered family F. Roughly speaking, we show that the easy cases one can think of (e.g. when edgeless graphs of the right sizes are in F, or if F contains only cliques) are the only families giving rise to a polynomial problem: all others are NP-complete. We then investigate the parameterized complexity of the problem and give similar sufficient conditions on F that give rise to $W[1]$-hard, $W[2]$-hard or FPT problems when the parameter is the size of the solution.

5.3. Modeling the flexibility of macro-molecules

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

5.3.1. Characterizing molecular flexibility by combining lRMSD measures

Participants: F. Cazals, R. Tetley.

The root mean square deviation (RMSD) and the least RMSD are two widely used similarity measures in structural bioinformatics. Yet, they stem from global comparisons, possibly obliterating locally conserved motifs. We correct these limitations with the so-called combined RMSD [26], which mixes independent lRMSD measures, each computed with its own rigid motion. The combined RMSD is relevant in two main scenarios, namely to compare (quaternary) structures based on motifs defined from the sequence (domains, SSE), and to compare structures based on structural motifs yielded by local structural alignment methods. We illustrate the benefits of combined RMSD over the usual RMSD on three problems, namely (i) the assignment of quaternary structures for hemoglobin (scenario #1), (ii) the calculation of structural phylogenies (case study: class II fusion proteins; scenario #1), and (iii) the analysis of conformational changes based on combined RMSD of rigid structural motifs (case study: one class II fusion protein; scenario #2). Using these, we argue that the combined RMSD is a tool a choice to perform positive and negative discrimination of degree of freedom, with applications to the design of move sets and collective coordinates. Combined RMSD are available within the Structural Bioinformatics Library (http: //sbl.inria.fr).

5.3.2. Multiscale analysis of structurally conserved motifs

Participants: F. Cazals, R. Tetley.

This work [25] develops a generic framework to perform a multiscale structural analysis of two structures (homologous proteins, conformations) undergoing conformational changes. Practically, given a seed structural alignment, we identify structural motifs with a hierarchical structure, characterized by three unique properties. First, the hierarchical structure sheds light on the trade-off between size and flexibility. Second, motifs can be combined to perform an overall comparison of the input structures in terms of combined RMSD, an improvement over the classical least RMSD. Third, motifs can be used to seed iterative aligners, and to design hybrid sequence-structure profile HMM characterizing protein families. From the methods standpoint, our framework is reminiscent from the bootstrap and combines concepts from rigidity analysis (distance difference matrices), graph theory, computational geometry (space filling diagrams), and topology (topological persistence). On challenging cases (class II fusion proteins, flexible molecules) we illustrate the ability of our tools to localize conformational changes, shedding light of commonalities of structures which would otherwise appear as radically different. Our tools are available within the Structural Bioinformatics Library (http://sbl.inria.fr). We anticipate that they will be of interest to perform structural comparisons at large, and for remote homology detection.

5.3.3. Hybrid sequence-structure based HMM models leverage the identification of homologous proteins: the example of class II fusion proteins

Participants: F. Cazals, R. Tetley.

In collaboration with P. Guardado-Calvo, J. Fedry, and F. Rey (Inst. Pasteur Paris, France).

In [27], we present a sequence-structure based method characterizing a set of functionally related proteins exhibiting low sequence identity and loose structural conservation. Given a (small) set of structures, our method consists of three main steps. First, pairwise structural alignments are combined with multi-scale geometric analysis to produce structural motifs i.e. regions structurally more conserved than the whole structures. Second, the sub-sequences of the motifs are used to build profile hidden Markov models (HMM) biased towards the structurally conserved regions. Third, these HMM are used to retrieve from UniProtKB proteins harboring signatures compatible with the function studied, in a bootstrap fashion. We apply these hybrid HMM to investigate two questions related to class II fusion proteins, an especially challenging class since known structures exhibit low sequence identity (less than 15%) and loose structural similarity (of the order of 15A in lRMSD). In a first step, we compare the performances of our hybrid HMM against those of sequence based HMM. Using various learning sets, we show that both classes of HMM retrieve unique species. The number of unique species reported by both classes of methods are comparable, stressing the novelty brought by our hybrid models. In a second step, we use our models to identify 17 plausible HAP2-GSC1 candidate sequences in 10 different drosophila melanogaster species. These models are not identified by the PFAM family HAP2-GCS1 (PF10699), stressing the ability of our structural motifs to capture signals more subtle than whole Pfam domains. In a more general setting, our method should be of interest for all cases functional families with low sequence identity and loose structural conservation. Our software tools are available from the FunChaT package of the Structural Bioinormatics Library (http://sbl.inria.fr).

5.3.4. Hamiltonian Monte Carlo with boundary reflections, and application to polytope volume calculations

Participants: F. Cazals, A. Chevallier.

In collaboration with S. Pion (Auctus, Inria Bordeaux).

This paper [23] studies HMC with reflections on the boundary of a domain, providing an enhanced alternative to Hit-and-run (HAR) to sample a target distribution in a bounded domain. We make three contributions. First, we provide a convergence bound, paving the way to more precise mixing time analysis. Second, we present a robust implementation based on multi-precision arithmetic – a mandatory ingredient to guarantee exact predicates and robust constructions. Third, we use our HMC random walk to perform polytope volume calculations, using it as an alternative to HAR within the volume algorithm by Cousins and Vempala. The tests, conducted up to dimension 50, show that the HMC RW outperforms HAR.

5.3.5. Wang-Landau Algorithm: an adapted random walk to boost convergence

Participants: F. Cazals, A. Chevallier.

The Wang-Landau (WL) algorithm is a recently developed stochastic algorithm computing densities of states of a physical system. Since its inception, it has been used on a variety of (bio-)physical systems, and in selected cases, its convergence has been proved. The convergence speed of the algorithm is tightly tied to the connectivity properties of the underlying random walk. As such, we propose in [22] an efficient random walk that uses geometrical information to circumvent the following inherent difficulties: avoiding overstepping strata, toning down concentration phenomena in high-dimensional spaces, and accommodating multidimensional distribution. Experiments on various models stress the importance of these improvements to make WL effective in challenging cases. Altogether, these improvements make it possible to compute density of states for regions of the phase space of small biomolecules.

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. A Sequential non-parametric multivariate two-sample test

Participant: F. Cazals.

In collaboration with A. Lhéritier (Amadeus, France).

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. Classically, this statistic is computed on the whole dataset, or is computed on a subset of the dataset by a function trained on its complement. We consider methods in a third tier [15], so as to deal with large (possibly infinite) datasets, and to automatically determine the most relevant scales to work at, making two contributions. First, we develop a generic sequential nonparametric testing framework, in which the sample size need not be fixed in advance. This makes our test a truly sequential nonparametric multivariate two-sample test. Under information theoretic conditions qualifying the difference between the tested distributions, consistency of the two-sample test is established. Second, we instantiate our framework using nearest neighbor regressors, and show how the power of the resulting two-sample test can be improved using Bayesian mixtures and switch distributions. This combination of techniques yields automatic scale selection, and experiments performed on challenging datasets show that our sequential tests exhibit comparable performances to those of state-of-the-art nonsequential tests.

5.4.2. Comparing two clusterings using matchings between clusters of clusters

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

In collaboration with R. Watrigant (University Lyon I, Laboratoire de l'Informatique du Parallélisme, France).

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. 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 and APX-hardness 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).

5.4.3. How long does it take for all users in a social network to choose their communities?

Participant: D. Mazauric.

In collaboration with J.-C. Bermond (Inria/I3S project-team Coati) and A. Chaintreau (Columbia University in the city of New York) and G. Ducoffe (National Institute for Research and Development in Informatics and Research Institute of the University of Bucharest).

We consider a community formation problem in social networks, where the users are either friends or enemies. The users are partitioned into conflict-free groups (*i.e.*, independent sets in the *conflict graph* $G^- = (V, E)$ that represents the enmities between users). The dynamics goes on as long as there exists any set of at most k users, k being any fixed parameter, that can change their current groups in the partition *simultaneously*, in such a way that they all strictly increase their utilities (number of friends *i.e.*, the cardinality of their respective groups minus one). Previously, the best-known upper-bounds on the maximum time of convergence were $\mathcal{O}(|V|\alpha(G^-))$ for $k \leq 2$ and $\mathcal{O}(|V|^3)$ for $k = 3$, with $\alpha(G^-)$ being the independence number of G^-. Our first contribution in this paper consists in reinterpreting the initial problem as the study of a dominance ordering over the vectors of integer partitions. With this approach, we obtain for $k \leq 2$ the tight upper-bound $\mathcal{O}(|V| \min\{\alpha(G^-), \sqrt{|V|}\})$ and, when G^- is the empty graph, the exact value of order $\frac{(2|V|)^{3/2}}{3}$. The time of convergence, for any fixed $k \geq 4$, was conjectured to be polynomial [35], [41]. In this paper we disprove this. Specifically, we prove that for any $k \geq 4$, the maximum time of convergence is an $\Omega(|V|^{\Theta(\log |V|)})$.

See [19] for details.

5.4.4. Sequential metric dimension

Participant: D. Mazauric.

In collaboration with J. Bensmail (I3S, Inria/I3S project-team Coati) and F. Mc Inerney (Inria/I3S project-team Coati) and N. Nisse (Inria, Inria/I3S project-team Coati) and S. Pérennes (CNRS, Inria/I3S project-team Coati).

In the localization game, introduced by Seager in 2013, an invisible and immobile target is hidden at some vertex of a graph G. At every step, one vertex v of G can be probed which results in the knowledge of the distance between v and the secret location of the target. The objective of the game is to minimize the number of steps needed to locate the target whatever be its location.

We address the generalization of this game where $k \geq 1$ vertices can be probed at every step. Our game also generalizes the notion of the *metric dimension* of a graph. Precisely, given a graph G and two integers $k, \ell \geq 1$, the *localization* problem asks whether there exists a strategy to locate a target hidden in G in at most ℓ steps and probing at most k vertices per step. We first show that, in general, this problem is NP-complete for every fixed $k \geq 1$ (resp., $\ell \geq 1$). We then focus on the class of trees. On the negative side, we prove that the localization problem is NP-complete in trees when k and ℓ are part of the input. On the positive side, we design a $(+1)$-approximation for the problem in n-node trees, *i.e.*, an algorithm that computes in time $O(n \log n)$ (independent of k) a strategy to locate the target in at most one more step than an optimal strategy. This algorithm can be used to solve the localization problem in trees in polynomial time if k is fixed.

We also consider some of these questions in the context where, upon probing the vertices, the relative distances to the target are retrieved. This variant of the problem generalizes the notion of the *centroidal dimension* of a graph.

See [17], [18], [21] for details.

6. Partnerships and Cooperations

6.1. International Research Visitors

6.1.1. Visits of International Scientists

6.1.1.1. Internships

- Internship of Maria Guramare, Harvard University, Cambridge, Massachusetts. Supervision: Frédéric Cazals and Dorian Mazauric. *Shortest Paths under Constraints Problem with Application for Structural Alignments.*
- Internship of Xuchun Zhang, École Polytechnique de l'Université Nice Sophia Antipolis, filière Mathématiques Appliquées et Modélisation, year 4 (Master 1). Supervision: Jean-Baptiste Caillau (Inria project-team McTao), Enzo Giusti (startup Oui!Greens), Dorian Mazauric, and Joanna Moulierac (Inria/I3S project-team Coati). *Problèmes d'affectations d'annonces dans un réseau anti gaspillage !*
- Project of Ruiqing Chang and Xuchun Zhang, École Polytechnique de l'Université Nice Sophia Antipolis, Filière Mathématiques Appliquées et Modélisation, year 4 (Master 1). Supervision: Jean-Baptiste Caillau (Inria project-team McTao), Enzo Giusti (startup Oui!Greens), Dorian Mazauric, and Joanna Moulierac (Inria/I3S project-team Coati). *Problèmes d'affectations d'annonces dans un réseau anti gaspillage !*

- Internship of Nguyen Thi Viet Ha, Master 2 Fundamental Computer Science, École Normale Supérieure de Lyon. Supervision: Frédéric Havet (Inria/I3S project-team Coati), Dorian Mazauric, and Rémi Watrigant (École Normale Supérieure de Lyon and Université Claude Bernard Lyon 1). *Graph Algorithms for low resolution model of large protein assemblies.*
- Internship of Timothée O'Donnell, Master 2 University Paris Saclay, Master bioinformatique. *Structural modeling of FMRP dimers in solution*. Supervision: F. Cazals.

7. Dissemination

7.1. Promoting Scientific Activities

7.1.1. Scientific Events Organisation

7.1.1.1. Member of the Organizing Committees

– Frédéric Cazals was member of the advisory board of:

- *Algorithms in Structural Bio-informatics*. The 2018/2019 edition (January 2019, CIRM, Marseille) focuses on RNA bioinformatics. See https://algosb2019.sciencesconf.org/.

7.1.2. Scientific Events Selection

7.1.2.1. Member of the 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) / Protein Interactions & Molecular Networks
- IEEE International Conference on BioInformatics and BioEngineering

7.1.3. Journal

7.1.3.1. Reviewer - Reviewing Activities

– Frédéric Cazals reviewed for the following journals:

- Journal of computational geometry
- PLOS Computational Biology

– Dorian Mazauric reviewed for the following journal and conference:

- Theoretical Computer Science
- 16th Workshop on Approximation and Online Algorithms (WAOA 2018)

7.1.4. Invited Talks

– Frédéric Cazals gave the following invited talks:

- *Energy landscapes: sampling, analysis, comparison*, RNA Kinetics days, Ecole polytechnique, October 2018.
- *Randomized algorithms for volume/density of states calculations in high-dimensional spaces*: Energy landscapes, Kalamata, Greece, September 2018;
- *Randomized algorithms for volume/density of states calculations in high-dimensional spaces*: Advances in Computational Statistical Physics, CIRM, France, September 2018.
- *Understanding scoring/energy landscapes: a tale of local minima and density of states*, Meet-U: when proteins meet each other, January 2018, Paris.

7.1.5. Leadership within the Scientific Community

– Frédéric Cazals:

- 2010-.... Member of the steering committee of the *GDR Bioinformatique Moléculaire*, 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 - Méditerranée.
- 2018-.... Member of the *bureau du comité des équipes projets*.

– Dorian Mazauric:

- 2016-2019. Member of the *Comité de Centre*, Inria Sophia Antipolis - Méditerranée.
- 2018-.... Member of the *Commission de Développement Technologique*, 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)
- Master : Dorian Mazauric, Algorithmique et Complexité, 36 h TD, niveau M1, École Polytechnique de l'Université Nice Sophia Antipolis, filière Sciences Informatiques, France.

7.2.2. Supervision

- **PhD:** Romain Tetley, *Mixed sequence-structure based analysis of proteins, with applications to functional annotations*, defended on the 21/11/2018. Université Côte d'Azur.
- **PhD in progress, 4th year:** Augustin Chevallier, *Random walks for estimating the volume of convex bodies and densities of states in high dimensional spaces*, defense scheduled in February 2019. Université Côte d'Azur.
- **PhD in progress, 2nd year:** Denys Bulavka, *Modeling macro-molecular motions*. Université Côte d'Azur. Under the supervision of Frédéric Cazals.
- **PhD in progress, 2nd year:** Méliné Simsir, *Modeling drug efflux by Patched*. Université Côte d'Azur. Thesis co-supervised by Frédéric Cazals and Isabelle Mus-Veteau, IPMC/CNRS.
- **PhD in progress, 1st year:** Timothée O'Donnel, *Modeling the influenza polymerase*. Université Côte d'Azur. Thesis co-supervised by Frédéric Cazals and Bernard Delmas, INRA Jouy-en-Josas.
- **PhD in progress, 1st year** : Thi Viet Ha Nguyen, Graph Algorithms techniques for (low and high) resolution models of large protein assemblies, Frédéric Havet (Inria/I3S project-team Coati) and Dorian Mazauric.

7.2.3. Juries

– Frédéric Cazals:

- Hugo Schweke, Paris-Saclay University, December 2018. Rapporteur on the PhD thesis *Développement d'une méthode* in silico *pour caractériser le potentiel d'interaction des surfaces protéiques dans un environnement encombré*. Advisors: Marie-Hélène Mucchielli-Giorgi and Anne Lopes.
- Julien Ogor, ENSTA Bretagne, May 2018. Rapporteur on the PhD thesis *Design of algorithms for the automatic characterization of marine dune morphology and dynamics*. Advisor: B. Zerr.
- Rodrigo Dorantes-Gilardi, University of Lyon, April 2018. Rapporteur on the PhD thesis *Bio-Mathematical aspects of the plasticity of proteins*. Advisors: L. Vuillon and C. Lesieur.

– Dorian Mazauric:

- Romain Tetley, Université Côte d'Azur, Novembre 2018. PhD thesis *Mixed sequence-structure based analysis of proteins, with applications to functional annotations*. Advisor: Frédéric Cazals.

7.3. Popularization

This part mainly concerns Dorian Mazauric.

7.3.1. Internal or external Inria responsibilities

- Member of Mastic Commission (Médiation et Animation scientifiques Inria Sophia Antipolis - Méditerranée).
- Coordinator of the popularization project GALEJADE (Graphes et ALgorithmes : Ensemble de Jeux À Destination des Écoliers (mais pas que)) founded by Inria, Fondation Blaise Pascal, and Université Côte d'Azur. See https://galejade.inria.fr.
- Coordinator of the internships for undergraduates of middle school (niveau collège, troisième) at Inria Sophia Antipolis - Méditerranée (12 interns during one week).

7.3.2. Articles and contents

Frédéric Cazals published the following opinion article:

- *Recherche et développement : les entreprises françaises n'ont pas de vision*, Le Monde, April 2018. See https://www.lemonde.fr/sciences/article/2018/04/20/r-d-les-entreprises-francaises-n-ont-pas-de-vision_5288104_1650684.html.

Dorian Mazauric published online contents and posters. See https://galejade.inria.fr.

7.3.3. Education

- Trainings for 100 future teachers at ÉSPÉ (École SupÉrieure du Professorat et de l'Éducation) of Académie de Nice.
- Two trainings for 60 teachers of Cycle 3 (Le Cannet).
- Trainings for 20 teachers at numeric culture week-end organised by Class'Code MED.

7.3.4. Interventions

- National events:
  - Fête de la Science : Village des Sciences de Vinon-sur-Verdon, Juan-les-Pins, Villeneuve-Loubet et Mouans Sartoux : *La magie des graphes et du binaire, Algorithmes grandeur nature et jeux combinatoires*.
  - Semaine des maths : Conferences and activities at Centre International de Valbonne. With Christophe Godin. *Réfléchir pour Calculer ou Calculer pour Réfléchir.*
  - Conferences and activities at salon Code & Play 2018. *Graphes et algorithmes ? Jeux grandeur nature : algorithme de plus court chemin, algorithme de tri avec des cerceaux et des lattes en plastique – La magie des graphes et du binaire : tours de magie.*
- In educational institutions:
  - Two trainings for 60 teachers of Cycle 3 (Le Cannet).
  - Trainings for 100 future teachers at ÉSPÉ (École SupÉrieure du Professorat et de l'Éducation) of Académie de Nice.
  - High school: Conferences at Centre International de Valbonne. *Pas besoin de réfléchir, les ordinateurs calculent tellement vite ? Théorie des graphes et algorithmique.*
  - Middle school: Conferences at collège Alphonse Daudet of Nice and conferences at collège Jules Verne of Cagnes-sur-Mer. *La magie des graphes et du binaire.*
  - Primary: Conferences at École élémentaire of Tourrettes-sur-Loup. With Florence Barbara.

- Welcoming of schoolchildren or the general public in an Inria center:
  - MathC2+ internship: Activity for 40 students (high school). With Maria Guramare. *Algorithmes grandeur nature pour le calcul du plus court chemin et pour trier.*
  - Open days of Inria Sophia Antipolis - Méditerranée: *La magie des graphes, des algorithmes et du binaire*.
  - Presentation for twelve interns of middle school (niveau collège, troisième) by Frédéric Cazals.

7.3.5. Creation of media or tools for science outreach

- Creation of the website of the popularization project GALEJADE (Graphes et ALgorithmes : Ensemble de Jeux À Destination des Écoliers (mais pas que)) founded by Inria, Fondation Blaise Pascal, and Université Côte d'Azur. See https://galejade.inria.fr.
- Development of wooden objects for the dissemination of the scientific culture: wooden plateau for graph algorithms and convex hull, chocolate bar game made by 3D printers, kakemonos... See https://galejade.inria.fr/francais-pret-de-materiel/.

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^{o} 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. *The Structural Bioinformatics Library: modeling in biomolecular science and beyond*, in "Bioinformatics", 2017, vol. 7, n^{o} 33, pp. 1–8 [*DOI :* 10.1093/bioinformatics/btw752], http://sbl.inria.fr

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

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

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

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

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

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

[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^{o} 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] N. COHEN, F. HAVET, D. MAZAURIC, I. SAU, R. WATRIGANT. *Complexity dichotomies for the Minimum F -Overlay problem*, in "Journal of Discrete Algorithms", September 2018, vol. 52-53, pp. 133-142 [*DOI :* 10.1016/J.JDA.2018.11.010], https://hal.inria.fr/hal-01947563

[15] A. LHÉRITIER, F. CAZALS. *A Sequential Non-Parametric Multivariate Two-Sample Test*, in "IEEE Transactions on Information Theory", May 2018, vol. 64, n^{o} 5, pp. 3361-3370, https://hal.inria.fr/hal-01968190

[16] S. MAGADAN, L. JOUNEAU, M. PUELMA TOUZEL, S. MARILLET, W. CHARA, A. SIX, E. QUILLET, T. MORA, A. WALCZAK, F. CAZALS, O. SUNYER, S. FILLATREAU, P. BOUDINOT. *Origin of Public Memory B Cell Clones in Fish After Antiviral Vaccination*, in "Frontiers in Immunology", September 2018, vol. 9, https://hal.inria.fr/hal-01968155

International Conferences with Proceedings

[17] J. BENSMAIL, D. MAZAURIC, F. MC INERNEY, N. NISSE, S. PÉRENNES. *Localiser une cible dans un graphe*, in "ALGOTEL 2018 - 20èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications", Roscoff, France, May 2018, https://hal.inria.fr/hal-01774827

[18] J. BENSMAIL, D. MAZAURIC, F. MC INERNEY, N. NISSE, S. PÉRENNES. *Sequential Metric Dimension*, in "16th Workshop on Approximation and Online Algorithms (WAOA 2018)", Helsinki, Finland, August 2018, https://hal.inria.fr/hal-01883712

[19] J.-C. BERMOND, A. CHAINTREAU, G. DUCOFFE, D. MAZAURIC. *How long does it take for all users in a social network to choose their communities?*, in "9th International Conference on Fun with Algorithms (FUN 2018)", La Maddalena, Italy, 2018, https://hal.inria.fr/hal-01780627

[20] F. CAZALS, D. MAZAURIC, R. TETLEY, R. WATRIGANT. *Comparaison de deux clusterings par couplage entre clusters de clusters*, in "ALGOTEL 2018 - 20èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications", Roscoff, France, May 2018, https://hal.inria.fr/hal-01774440

Research Reports

[21] J. BENSMAIL, D. MAZAURIC, F. MC INERNEY, N. NISSE, S. PERENNES. *Sequential Metric Dimension*, Inria, 2018, https://hal.archives-ouvertes.fr/hal-01717629

[22] A. CHEVALLIER, F. CAZALS. *Wang-Landau Algorithm: an adapted random walk to boost convergence*, Inria Sophia Antipolis, France, November 2018, https://hal.archives-ouvertes.fr/hal-01919860

[23] A. CHEVALLIER, S. PION, F. CAZALS. *Hamiltonian Monte Carlo with boundary reflections, and application to polytope volume calculations*, Inria Sophia Antipolis, France, November 2018, n^{o} RR-9222, https://hal.archives-ouvertes.fr/hal-01919855

Other Publications

[24] J.-C. BERMOND, D. MAZAURIC, V. MISRA, P. NAIN. *Distributed Link Scheduling in Wireless Networks*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01977266

[25] F. CAZALS, R. TETLEY. *Multiscale analysis of structurally conserved motifs*, July 2018, working paper or preprint, https://hal.inria.fr/hal-01968176

[26] R. TETLEY, F. CAZALS. *Characterizing molecular flexibility by combining lRMSD measures*, July 2018, working paper or preprint, https://hal.inria.fr/hal-01968175

[27] R. TETLEY, F. CAZALS. *Hybrid sequence-structure based HMM models leverage the identification of homologous proteins: the example of class II fusion proteins*, July 2018, working paper or preprint, https://hal.inria.fr/hal-01968177

References in notes

[28] 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

[29] 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

[30] 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

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

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

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

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

[35] B. ESCOFFIER, L. GOURVÈS, J. MONNOT. *Strategic coloring of a graph*, in "Internet Mathematics", 2012, vol. 8, n^{o} 4, pp. 424–455

[36] 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

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

[38] 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

[39] 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

[40] 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

[41] J. KLEINBERG, K. LIGETT. *Information-sharing in social networks*, in "Games and Economic Behavior", 2013, vol. 82, pp. 702–716

[42] 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

[43] 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

[44] 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

[45] 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

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

[47] 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

[48] 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

[49] 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

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

[51] 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

[52] 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

[53] 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

[54] 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

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