INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team abs

Algorithms, Biology, Structure

Sophia Antipolis - Méditerranée

THEME COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Table of contents

1. Team

Research Scientist

Frédéric Cazals [DR2 Inria, HdR]
Julie Bernauer [CR2 Inria]

PhD Student

Sébastien Loriot [MESR monitor fellow; January - November 2008]
Tom Dreyfus [MESR monitor fellow]

Post-Doctoral Fellow

Sébastien Loriot [INRIA - Direction Scientifique; December 2008]

Administrative Assistant

Agnès Bessière [TR Inria, assistant of GEOMETRICA and ABS; until August 2008]

2. Overall Objectives

2.1. Introduction

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, 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 question 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 [44]. 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, while the order of magnitude of the number of genomes sequenced is one thousand, the Protein Data Bank contains (a mere) 45,000 structures. (Because one gene may yield a number of proteins through splicing, it is difficult to estimate the number of proteins from the number of genes. However, the latter is several orders of magnitudes beyond the former.) For these reasons, *molecular modeling* is expected to play a key role in investigating structural issues.

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 [42], [30] and later Connolly [26], 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) [33], 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, which is out of 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.

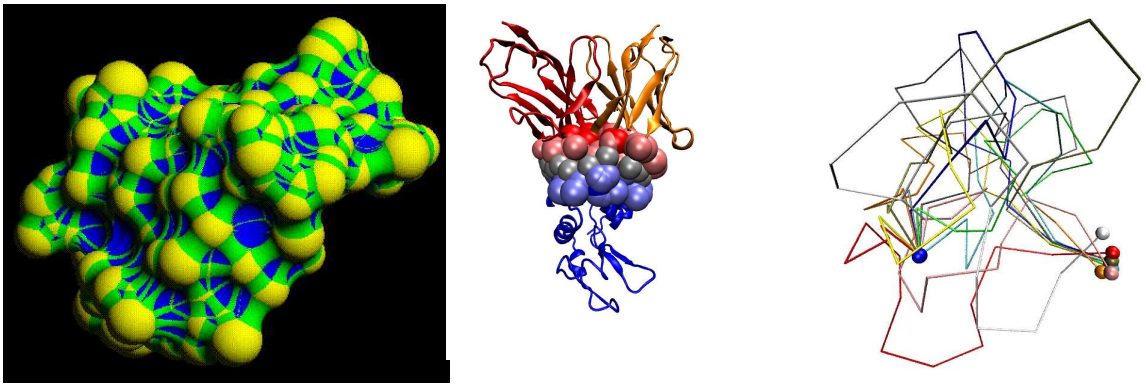

Figure 1. (a) Molecular surface (b) An antibody-antigen complex, with interface atoms computed as described in [9] (c)Conformations of a backbone loop

2.2. Highlights of the year

The ABS project-team was created in July 2008.

The first PhD defense, that of S. Loriot, occurred on 12/02/2008.

F. Cazals was awarded a PhD grant by the Ministry of Research from the pool of 555 grants on *Thématiques prioritaires.*

3. Scientific Foundations

3.1. Introduction

The research conducted by ABS focuses on two main directions in Computational Structural Biology (CSB), each such direction calling for specific algorithmic developments. These directions are:
- Modeling interfaces and contacts,
- Modeling the flexibility of macro-molecules.

3.2. Modeling Interfaces and Contacts

Problems addressed. 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 [44]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [47]. Current investigations follow two routes. From the experimental perspective [29], directed mutagenesis allows one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [41]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [36].

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 [21], 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 [45], [32]. 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 [37]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

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

[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.

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 relate the geometry and the bio-chemistry at the interface level [9]. 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 [22]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [46], 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 [28].

A structural alphabet at the atomic level may be seen as an alphabet featuring for an atom of a given type all the conformations this atom may engage into, depending on its neighbors. One way to tackle this problem consists of extending the notions of molecular surfaces used so far, so as to encode multi-body relations between an atom and its neighbors [15]. In order to derive such alphabets, the following two strategies are obvious. On one hand, one may use an encoding of neighborhoods based on geometric constructions such as Voronoï 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.

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 [40]. 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 the Flexibility of Macro-molecules

Problems addressed. 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 and Molecular Dynamics simulations generate ensembles of conformations, called `conformers`. 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.

Methodological developments. At the side-chain level, the question of improving rotamer libraries is still of interest [27]. 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) [43]. 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 [39], to Morse theory [34] and to analysis of meta-stable states of time series [35] have been proposed.

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

Given a sampling on an energy landscape, a number of fundamental issues actually arise: how does the point cloud describes 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 noticed in passing such questions are probably related to modeling phase transitions in statistical physics where geometric and topological methods are being used [38].

From an algorithmic standpoint, such questions are reminiscent from *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 CAGD, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years [6]. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison [23]; the study of Morse-like constructions stemming from distance functions to points [31]; the analysis of topological invariants of the model and the samples, and their comparison [24], [25].

Last but not the least, gaining insight on such questions would also help to effectively select a reduced set of conformations best representing a larger number of conformations. This selection problem is indeed faced by flexible docking algorithms that need to maintain and/or update collections of conformers for the second stage of the *diffusion - conformer selection - induced fit* complex formation model.

4. Software

4.1. Web services

4.1.1. Modeling macro-molecular interfaces

Participant: Frédéric Cazals.

We recently proposed an interface model of (macro-)molecular interfaces based upon power diagrams [9]. The corresponding software, *Intervor*, has been made available to the community from the web site http://cgal.inria.fr/Intervor. Our publication appeared in the September 2006 issue of *Protein Science*, and the server has been used about 1000 times since then. To the best of our knowledge, this code is the only publicly available one for analyzing (Voronoi) interfaces in complexes.

4.1.2. Discrimination between crystallographic and biological protein-protein interactions

Participant: Julie Bernauer.

Following the recent strategy we developed to identify biological and crystallographic contacts using support vector machines [12], we made a web server available for that purpose. The DiMoVo server http://cgal.inria.fr/DiMoVo/ has been available from March 2008.

4.1.3. Protein-protein docking conformation evaluation

Participant: Julie Bernauer.

Available online in 2007, the VorScore server http://cgal.inria.fr/VorScore/ is now accessible from the team webservice area. It gives access to the scoring function evaluation published in 2007 [1].

4.2. CGAL and Ipe

Participant: Sébastien Loriot.

In collaboration with L. Rineau and S. Pion, GEOMETRICA. *Work started by Nicolas Carrez, summer intern, 2005. http://www.cgal.org*

CGAL is a C++ library of geometric algorithms initially developed within two European projects (project ESPRIT IV LTR CGAL December 97 - June 98, project ESPRIT IV LTR GALIA november 99 - august 00) by a consortium of eight research teams from the following institutes: Universiteit Utrecht, Max-Planck Institut Saarbrücken, INRIA Sophia Antipolis, ETH Zürich, Tel Aviv University, Freie Universität Berlin, Universität Halle, RISC Linz. The goal of CGAL is to make the solutions offered by the computational geometry community available to the industrial world and applied domains.

The IPE editor, see http://tclab.kaist.ac.kr/ipe, is a graphical editor which combines XFIG like facilities together with standard Computational Geometry algorithms. It is intensively used by the computational geometry community for making presentations as well as illustrating papers.

Based on the $2D$ algorithms present in the CGAL library, we developed in C++ a set of plugins, so as to make the following algorithms available from IPE: triangulations (Delaunay, constrained Delaunay, regular) as well as their duals, a convex hull algorithm, polygon partitioning algorithms, polygon offset, arrangements of linear and degree two primitives. These plugins are available under the Open Source LGPL license, and are subject to the constraints of the underlying CGAL packages. They can be downloaded from http://cgal-ipelets.gforge.inria.fr.

5. New Results

5.1. Modeling Interfaces and Contacts

5.1.1. Shelling the Voronoi interface of Protein-protein Complexes Predicts Residue Activity and Conservation

Participants: Benjamin Bouvier, Raik Gruenberg, Michael Nilgès, Frédéric Cazals.

B. Bouvier is with Institut de Biologie et de Chimie des Protéines, CNRS/Lyon Univ., France; R. Grünberg is with EMBL-CRG Systems Biology Unit, Barcelona, Spain; Nilges is with Unité de Bioinformatique Structurale, Institut Pasteur Paris, France.

The accurate description and analysis of protein-protein interfaces remains a challenging task. Traditional definitions, based on atomic contacts or changes in solvent accessibility, tend to over- or underpredict the interface itself and cannot discriminate active from less relevant parts. This paper [19] introduces a fast, parameter-free and purely geometric definition of protein interfaces and introduce the shelling order of Voronoi facets as a novel measure for an atom's depth inside the interface. Our analysis of 54 protein-protein complexes reveals a strong correlation between Voronoi Shelling Order (VSO) and water dynamics. High Voronoi Shelling Order coincides with residues that were found shielded from bulk water fluctuations in a recent molecular dynamics study. Yet, VSO predicts such "dry" residues without consideration of forcefields or dynamics at dramatically reduced cost. More central interface positions are often also increasingly enriched for hydrophobic residues. Yet, this hydrophobic centering is not universal and does not mirror the far stronger geometric bias of water fluxes. The seemingly complex water dynamics at protein interfaces appears thus largely controlled by geometry. Sequence analysis supports the functional relevance of both dry residues and residues with high VSO, both of which tend to be more conserved. Upon closer inspection, the spatial distribution of conservation argues against the arbitrary dissection into core or rim and thus refines previous results. Voronoi Shelling Order reveals clear geometric patterns in protein interface composition, function and dynamics and facilitates the comparative analysis of protein-protein interactions.

5.1.2. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions

Keywords: *Crystallographic contacts*, *Protein interactions*, *Scoring function;Voronoi diagrams*, *Support Vector Machines*.

Participants: Julie Bernauer, Ranjit Prasad Bahadur, Francis Rodier, Joël Janin, Anne Poupon.

R. Bahadur is Postdoctoral Scholar at the Jacobs University of Bremen, Germany; F. Rodier is retired from LEBS in CNRS Gif-sur-Yvette, France; J. Janin is Emeritus Professor at the Université Paris-Sud, Orsay, France; A. Poupon is in the Physiologie de la Reproduction et des Comportements lab, INRA Tours.

Knowledge of the oligomeric state of a protein is often essential for understanding its function and mechanism. Within a protein crystal, each protein monomer is in contact with many others, forming many small interfaces and a few larger ones that are biologically significant if the protein is a homodimer in solution, but not if the protein is monomeric. Telling such "crystal dimers" from real ones remains a difficult task.

It has already been demonstrated that the interfaces of native and non-native protein-protein complexes can be distinguished using a combination of parameters computed with a method on the Voronoi tessellation. We show in our study [12] that the same parameters highlight significant differences between the interfaces of biological and crystal dimers. Using these parameters as descriptors in machine learning methods leads to accurate classification of specific and non-specific protein-protein interfaces.

5.2. Modeling the flexibility of macro-molecules

5.2.1. Characterizing and Selecting Diverse Conformer Ensembles for Docking

Participants: Sébastien Loriot, Karine Bastard, Sushant Sachdeva, Chantal Prévost.

S. Sachdeva is currently PhD student at Princeton University; K. Bastard is with Biotechnologie-Biocatalyse-Biorégulation, Université de Nantes - CNRS, France; C. Prévot is with Institut de Biologie Physico-Chimique, Paris, France.

To address challenging flexible docking problems, a number of docking algorithms pre-generate large collections of candidate conformers. To further remove the redundancy from such ensembles, a central question in this context is the following one: report a selection of conformers maximizing some geometric diversity criterion. In this context, this paper [20] makes three contributions.

First, we tackle this problem resorting to geometric optimization so as to report selections maximizing the molecular volume or molecular surface area (MSA) of the selection. Greedy strategies are developed, together with approximation bounds.

Second, to assess the efficacy of our algorithms, we investigate two conformer ensembles corresponding to a flexible loop of four protein complexes. By focusing on the MSA of the selection, we show that our strategy matches the MSA of standard selection methods, but resorting to a number of conformers between one and two orders of magnitude smaller. This observation is qualitatively explained using the Betti numbers of the union of balls of the selection.

Finally, we replace the conformer selection problem in the context of multiple-copy flexible docking. On the systems above, we show that using the loops selected by our strategy can significantly improve the result of the docking process.

5.2.2. Exploring Point Cloud with Applications to Collective Coordinates

Participants: Frédéric Cazals, Frédéric Chazal, Joachim Giesen.

F. Chazal is with INRIA Saclay - Geometrica; J. Giesen is Professor at Jena University.

Life sciences, engineering, or telecommunications provide numerous systems whose description requires a large number of variables. Developing insights into such systems, forecasting their evolution, or monitoring them is often based on the inference of correlations between these variables. Given a collection of points describing states of the system, questions such as inferring the effective number of independent parameters of the system (its intrinsic dimensionality) and the way these are coupled are paramount to develop models. In this context, this paper [18] makes two contributions.

First, we review recent work on spectral techniques to organize point clouds in Euclidean space, with emphasis on the main difficulties faced. Second, after a careful presentation of the bio-physical context, we present applications of dimensionality reduction techniques to a core problem in structural biology, namely protein folding.

Both from the computer science and the structural biology perspective, we expect this survey to shed new light on the importance of *non linear computational geometry* in geometric data analysis in general, and for protein folding in particular.

5.3. Algorithmic foundations

5.3.1. Robust Construction of the Three-dimensional Flow Complex

Participants: Frédéric Cazals, Aaditya Parameswaran, Sylvain Pion.

S. Pion is with INRIA Geometrica (Sophia-Antipolis), and A. Parameswaran is currently PhD student at Stanford University.

The Delaunay triangulation and its dual the Voronoi diagram are ubiquitous geometric complexes. From a topological standpoint, the connection has recently been made between these cell complexes and the Morse theory of distance functions. In particular, in the generic setting, algorithms have been proposed to compute the flow complex —the stable and unstable manifolds associated to the critical points of the distance function to a point set. As algorithms ignoring degenerate cases and numerical issues are bound to fail on general inputs, this paper [17] develops the first complete and robust algorithm to compute the flow complex.

First, we present complete algorithms for the flow operator, unraveling a delicate interplay between the degenerate cases of Delaunay and those which are flow specific. Second, we sketch how the flow operator unifies the construction of stable and unstable manifolds. Third, we discuss numerical issues related to predicates on cascaded constructions. Finally, we report experimental results with CGAL's filtered kernel, showing that the construction of the flow complex incurs a small overhead w.r.t. the Delaunay triangulation when moderate cascading occurs. These observations provide important insights on the relevance of the flow complex for (surface) reconstruction and medial axis approximation, and should foster flow complex based algorithms.

In a broader perspective and to the best of our knowledge, this paper is the first one reporting on the effective implementation of a geometric algorithm featuring cascading.

5.3.2. Reporting Maximal Cliques

Participants: Frédéric Cazals, Chinmay Karande.

C. Karande is currently PhD student at Georgia Tech.

Reporting the maximal cliques of a graph is a fundamental problem arising in many areas. This note [14] bridges the gap between three papers addressing this problem: the original paper of Bron-Kerbosh (Comm. of the ACM, 1973), and two papers recently published in *Theoretical Computer Science*, namely that of Tomita et al. (*Theoretical Computer Science* 363, 2006), and that of Koch (*Theoretical Computer Science* 250, 2001). In particular, we show that the strategy of Tomita et al. is a simple modification of the Bron-Kerbosch algorithm, based on an (un-exploited) observation raised in Koch's paper.

6. Other Grants and Activities

6.1. International initiatives

6.1.1. Project France-Stanford Center for Interdisciplinary Studies

Participants: Julie Bernauer, Frédéric Cazals, Sébastien Loriot.

The France-Stanford Center for Interdisciplinary Studies is funding a two-year project (2007-08) entitled *Developments of Geometric Methods and Algorithms for the study of macro molecular assemblies*. The PIs are F. Cazals (INRIA) and M. Levitt (chair of the Structural Biology Dpt, Stanford University). The goal of the project is to make a stride towards improved multi-scale modeling of large protein complexes.

7. Dissemination

7.1. Animation of the scientific community

7.1.1. Conference program committees

– Frédéric Cazals was member of the paper committees of the Eurographics Symposium on Geometry Processing'08, of the ACM Symposium on Solid and Physical Modeling'08, of the International Conference on Pattern Recognition in Bioinformatics'08, of the International Symposium on 3D Data Processing, Visualization, and Transmission'08, and of the Symposium on Point-Based Graphics'08.

7.1.2. Ph.D. thesis and HDR committees

– Frédéric Cazals was reviewer for the Ph.D. thesis of Christine Martin (Univ. Orsay).

7.1.3. Recruitment committees

– Frédéric Cazals was a member of two committes at Université d'Evry-Val-d'Essonne, for the recruitment of one *Professor* and one *Maître de conferences* in CS/Bioinformatics.

7.1.4. Structural Biology and Modelling Workgroup

Modelling in structural biology is a topic of interest for a number of groups around Sophia-Antipolis and Nice, both in academia and (CNRS, université de Nice-Sophia-Antipolis, INRA, INRIA), and industry (in particular Galderma, one of the worldwide leading dermatology companies). Researchers from these organizations meet half a day once every trimester, to attend two talks on topics of general interest in this realm. The organization of these meetings for the academic year 2008-09 is handled by ABS.

7.1.5. WWW server

http://www-sop.inria.fr/abs/
The ABS web server is up and running since July 2008.

7.2. Teaching

7.2.1. Teaching at universities

- Master Bioinformatique et Biostatistiques (BIBS), Orsay University; Algorithmic Problems in Computational Structural Biology; F. Cazals (12h), J. Janin (6h), C. Robert (6h).
- Cursus Ecole Normale Supérieure de Lyon à Sophia-Antipolis; Topics in Structural Biology; F. Cazals (16h), J. Bernauer (8h).
- AgroParisTech, Paris, MAP3 (module d'approfondissement) Ingénierie des protéines, cursus ingénieur agronome, deuxième année; Introduction à la bioinformatique; J.Bernauer (3h).
- Polytech'Nice-Sophia, troisième année, Option Bio-Informatique et Modélisation pour la Biologie; Biogeometry; J.Bernauer (3h).

7.2.2. Internships

Internship proposals can be seen on the web from the Positions section at http://www-sop.inria.fr/abs/
– Aaditya Ramdas, *Landmark based dimensionality reduction*, IIT Bombay.

– Nicolas Bonifas, *Comparing Voronoi interfaces of protein-protein complexes*, ENS Lyon.

7.2.3. Ongoing Ph.D. theses

– Tom Dreyfus, *Modeling large macro-molecular assemblies*, université de Nice Sophia-Antipolis.

– Sébastien Loriot, *Arrangements de Cercles sur une Sphère: Algorithmes et Applications aux Modèles Moléculaires Representés par une Union de Boules*, université de Bourgogne. Defended on 12/02 in front of the following committee: R. Lavery (Univ. Lyon - CNRS; rapporteur); J. Snoeyink (Univ. North Carolina, rapporteur); John Maddock (Ecole Polytechnique Fédérale de Lausanne; examinateur); F. Chazal (INRIA Saclay - Geometrica; co-advisor); F. Cazals (INRIA Sophia-Antipolis - ABS, co-advisor).

7.3. Participation to conferences, seminars, invitations

7.3.1. Invited talks

Members of the project have presented their published articles at conferences. The reader can refer to the bibliography to obtain the corresponding list. We list below all other talks given in seminars or summer schools.

– *Modeling protein - protein interactions: the geometry of active residues and the selection of conformers*; IRISA, Rennes, 05/08. F.Cazals.

– *Protein folding: energy landscapes, spectral analysis, and Morse theory*; Journées de Dynamique Non Linéaire, Marseille; 01/08. F.Cazals.

– *Describing protein-protein and atomic environments: a geometric perspective*; (i) Séminaire Mathématiques Appliquées à la Génomique: Modèles et Algorithmes, Marseille, 01/08; (ii) Université de Nice, séminaire du Dpt de Mathématiques, 01/08. F.Cazals.

– *Protein - Protein interactions: towards improved predictions*; Sanofi - Aventis research seminar, 12/08.

– *Machine learning and protein docking scoring functions*; EMBO Workshop "Docking Predictions of Protein-Protein Interaction", Barcelona, 14-17 October 08. J.Bernauer.

7.3.2. The ABS seminar

The ABS seminar series started in 2008 and featured presentations from the following visiting scientists:
– M. Levitt, Stanford University, USA;
– S. Dokudovskaya, Inst. Jacques Monod, Paris, France;
– H-X. Zhou, Florida State University, USA;
– S. Flores, Stanford University, USA.

7.3.3. Scientific visits

ABS has hosted the following scientists:
– Dahlia Weiss, visiting PhD student from Stanford University, from 01/30/08 to 06/16/08.

8. Bibliography

Major publications by the team in recent years

[1] J. BERNAUER, J. AZE, J. JANIN, A. POUPON. *A new protein-protein docking scoring function based on interface residue properties.*, in "Bioinformatics", vol. 23, n^{o} 5, 2007, p. 555-62, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17237048.

[2] J. BERNAUER, A. POUPON, J. AZE, J. JANIN. *A docking analysis of the statistical physics of protein-protein recognition.*, in "Phys Biol", vol. 2, n^{o} 2, 2005, p. S17-23, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16204845.

[3] J.-D. BOISSONNAT, F. CAZALS. *Smooth Surface Reconstruction via Natural Neighbour Interpolation of Distance Functions*, in "Comp. Geometry Theory and Applications", 2002, p. 185–203.

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

[5] F. CAZALS, F. CHAZAL, T. LEWINER. *Molecular shape analysis based upon the Morse-Smale complex and the Connolly function*, in "ACM Symposium on Computational Geometry", 2003.

[6] F. CAZALS, J. GIESEN. *Delaunay Triangulation Based Surface Reconstruction*, in "Effective Computational Geometry for curves and surfaces", D.-D. BOISSONNAT, M. TEILLAUD (editors), Springer-Verlag, Mathematics and Visualization, 2006.

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

[8] F. CAZALS, M. POUGET. *Estimating Differential Quantities using Polynomial fitting of Osculating Jets*, in "Computer Aided Geometric Design", Conf. version: Symp. on Geometry Processing 2003, vol. 22, n^{o} 2, 2005, p. 121–146.

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

[10] S. GRAZIANI, J. BERNAUER, S. SKOULOUBRIS, M. GRAILLE, C. Z. ZHOU, C. MARCHAND, P. DECOTTIGNIES, H. VAN TILBEURGH, H. MYLLYKALLIO, U. LIEBL. *Catalytic mechanism and structure of viral flavin-dependent thymidylate synthase ThyX.*, in "J Biol Chem", vol. 281, n^{o} 33, 2006, p. 24048-57, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16707489.

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] S. LORIOT. *Arrangements de Cercles sur une Sphère: Algorithmes et Applications aux Modèles Moléculaires Representés par une Union de Boules*, Ph. D. Thesis, Université de Bourgogne, December 2008.

Articles in International Peer-Reviewed Journal

[12] J. BERNAUER, R. P. BAHADUR, F. RODIER, J. JANIN, A. POUPON. *DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.*, in "Bioinformatics", vol. 24, n^{o} 5, 2008, p. 652-8, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18204058.

[13] P. M. M. D. CASTRO, F. CAZALS, S. LORIOT, M. TEILLAUD. *Design of the CGAL Spherical Kernel and application to arrangements of circles on a sphere*, in "Computational Geometry: Theory and Applications", To appear, 2008.

[14] F. CAZALS, C. KARANDE. *A note on the problem of reporting maximal cliques*, in "Theoretical Computer Science", vol. 407, n^{o} 1–3, 2008, p. 564–568.

[15] F. CAZALS, S. LORIOT. *Computing the exact arrangement of circles on a sphere, with applications in structural biology*, in "Computational Geometry: Theory and Applications", To appear, 2008.

[16] F. CAZALS, M. POUGET. *Algorithm 889: Jet_fitting_3—A Generic C++ Package for Estimating the Differential Properties on Sampled Surfaces via Polynomial Fitting*, in "ACM Transactions on Mathematical Software", vol. 35, n^{o} 3, 2008.

International Peer-Reviewed Conference/Proceedings

[17] F. CAZALS, A. PARAMESWARAN, S. PION. *Robust construction of the three-dimensional flow complex*, in "ACM Symposium on Computational Geometry", 2008, p. 182–191.

Scientific Books (or Scientific Book chapters)

[18] F. CAZALS, F. CHAZAL, J. GIESEN. *Spectral Techniques to Explore Point Clouds in Euclidean Space, with Applications to the Inference of Collective Coordinates in Structural Biology*, in "Nonlinear Computational Geometry", I. EMIRIS, F. SOTTILE, T. THEOBALD (editors), To appear, The Institute of Mathematics and its Applications, 2008.

Research Reports

[19] B. BOUVIER, R. GRUNBERG, M. NILGES, F. CAZALS. *Shelling the Voronoi interface of protein-protein complexes predicts residue activity and conservation*, Submitted, Technical report, n^{o} 6415, INRIA, 2008, http://hal.inria.fr/inria-00206173/en/.

[20] S. LORIOT, S. SACHDEVA, K. BASTARD, C. PREVOST, F. CAZALS. *On the Characterization and Selection of Diverse Conformational Ensembles*, Submitted, Technical report, n^{o} 6503, INRIA, 2008, http://hal.inria.fr/inria-00252046/en/.

References in notes

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

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

[23] F. CHAZAL, D. COHEN-STEINER, A. LIEUTIER. *A Sampling Theory for Compacts in Euclidean Spaces*, in "ACM Symp. Comp. Geometry", 2006.

[24] F. CHAZAL, A. LIEUTIER. *Weak Feature Size and persistent homology : computing homology of solids in $\mathbb{R}^n$ from noisy data samples*, 2005.

[25] D. COHEN-STEINER, H. EDELSBRUNNER, J. HARER. *Stability of Persistence Diagrams*, in "ACM Symp. Comp. Geometry", 2005.

[26] M. L. CONNOLLY. *Analytical molecular surface calculation*, in "J. Appl. Crystallogr.", vol. 16, 1983.

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

[28] A. FERNÁNDEZ, R. BERRY. *Extent of Hydrogen-Bond Protection in Folded Proteins: A Constraint on Packing Architectures*, in "Biophysical Journal", vol. 83, 2002, p. 2475-2481.

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

[30] M. GERSTEIN, F. RICHARDS. *Protein geometry: volumes, areas, and distances*, in "The international tables for crystallography (Vol F, Chap. 22)", vol. F (Chapter 22.1.1), 2001.

[31] J. GIESEN, M. JOHN. *The Flow Complex: A Data Structure for Geometric Modeling*, in "ACM SODA", 2003.

[32] H. GOHLKE, G. KLEBE. *Statistical potentials and scoring functions applied to protein-ligand binding*, in "Curr. Op. Struct. Biol.", vol. 11, 2001, p. 231-235.

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

[34] V. K. KRIVOV, M. KARPLUS. *Hidden complexity of free energy surfaces for peptide (protein) folding*, in "PNAS", vol. 12, 2004.

[35] 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.

[36] I. MIHALEK, O. LICHTARGE. *On Itinerant Water Molecules and Detectability of Protein-Protein Int erfaces through Comparative Analysis of Homologues*, in "JMB", vol. 369, n^{o} 2, 2007.

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

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

[39] 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", vol. 67, n^{o} 4, 2007, p. 897–907.

[40] D. RAJAMANI, S. THIEL, S. VAJDA, C. CAMACHO. *Anchor residues in protein-protein interactions*, in "PNAS", vol. 101, 2004, p. 11287-11292.

[41] 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", vol. 102, n^{o} 1, 2005, p. 57-62, http://www.pnas.org/cgi/content/abstract/102/1/57.

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

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

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

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

[46] C. SUMMA, M. LEVITT, W. DEGRADO. *An atomic environment potential for use in protein structure prediction*, in "JMB", vol. 352, 2005.

[47] S. WODAK, J. JANIN. *Structural basis of macromolecular recognition*, in "Adv. in protein chemistry", vol. 61, 2003.