INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team abs

Algorithms, Biology, Structure

Sophia Antipolis - Méditerranée

Theme : Computational Biology and Bioinformatics

Table of contents

1. Team

Research Scientists

Frédéric Cazals [Team leader; DR2 Inria, HdR]

Julie Bernauer [CR2 Inria; Visiting the INRIA AMIB project-team until the 07/31/2010; member of AMIB project-team from the 08/01/2010.]

PhD Students

Tom Dreyfus [MESR monitor fellow]

Christine-Andrea Roth [INRIA CORDI-S fellow, from the 10/15/2010]

Post-Doctoral Fellow

Noël Malod-Dognin [From the 07/01/2010]

Administrative Assistant

Caroline French [Assistant of GEOMETRICA and ABS]

Others

Christine-Andrea Roth [Master intern from the MVA Cachan Master program; 04/15/2010 – 09/15/2010]

Ammad Ud-in [Master intern from the Computational Biology Master program; 03/01/2010 – 08/31/2010]

Palak Dalal [Summer intern from IIT Bombay; 05/01/2010 – 07/15/2010]

Achin Bansal [Summer intern from IIT Bombay; 05/01/2010 – 07/15/2010]

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 questions in structural biology. This goal actually subsumes two equally difficult challenges, which are *folding* —the process through which a protein adopts its $3d$ structure, and *docking* —the process through which two or several molecules assemble. Folding and docking are driven by non covalent interactions, and for complex systems, are actually inter-twined [45]. Apart from the bio-physical interests raised by these processes, two different application domains are concerned: in fundamental biology, one is primarily interested in understanding the machinery of the cell; in medicine, applications to drug design are developed.

Modeling in Computational Structural Biology. Acquiring structural data is not always possible: NMR is restricted to relatively small molecules; membrane proteins do not crystallize, etc. As a matter of fact, 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 [43], [31] and later Connolly [27], while defining molecular surfaces and developing algorithms to compute them, established the connexions between molecular modeling and geometric constructions. Remarkably, a number of difficult problems (e.g. additively weighted Voronoi diagrams) were touched upon in these early days.

The models developed in this vein are instrumental in investigating the interactions of molecules for which no structural data is available. But such models often fall short from providing complete answers, which we illustrate with the folding problem. On one hand, as the conformations of side-chains belong to discrete sets (the so-called rotamers or rotational isomers) [34], the number of distinct conformations of a poly-peptidic chain is exponential in the number of amino-acids. On the other hand, Nature folds proteins within time scales ranging from milliseconds to hours, 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.

2.2. Highlights

We achieved significant results on the problem of modeling macro-molecular complexes, at different scales.

Concerning the atomic-level modeling of binary complexes, we officially released `Intervor` , a program implementing structural descriptors improving the description of protein - protein interfaces [10], [2]. Thanks to the aforementioned Bioinformatics paper and the web-site http://cgal.inria.fr/abs/Intervor/, more than one thousand calculations have been run on our server over the past year, and the binary has been downloaded about 60 times in the same period. (The number of downloads is significant with respect to the size of the community into protein docking.)
In fact, we expect `Intervor` to become one of the standard tools to model protein interfaces, and we also look forward to specific developments for the important class of antibody-antigen complexes.

Concerning the intermediate resolution modeling of large protein assemblies, we made a stride [15] on the problem of making a precise assessment of *fuzzy / uncertain* models reconstructed from data integration [20]. In a nutshell, the reconstruction of assemblies involving hundreds of polypeptide chains is inherently challenging due to uncertainties on the various data involved in such a task. We proposed a framework allowing one to inherently model with uncertainties. The framework consists of replacing a possibly erroneous fixed

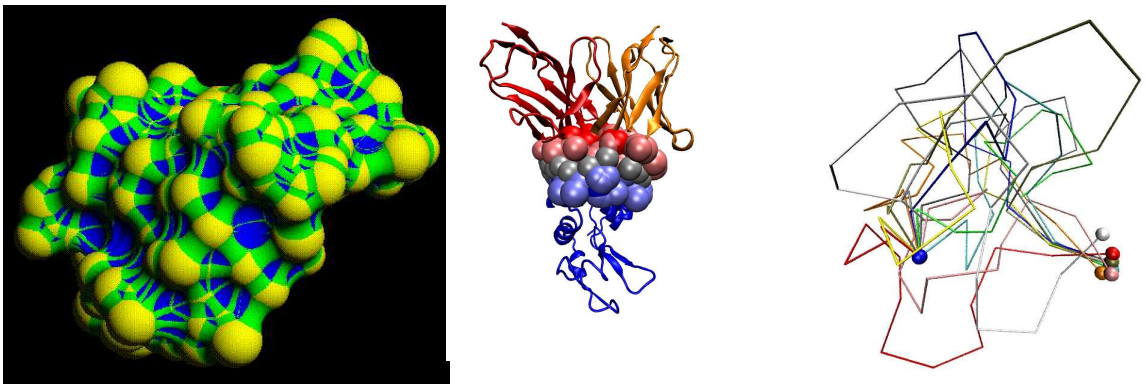

Figure 1. (a) Molecular surface (b) An antibody-antigen complex, with interface atoms computed as described in [10] (c)Conformations of a backbone loop

shape by a one-parameter family of shapes. This family consists of a finite collection of so-called *toleranced balls*, a toleranced ball being a ball whose radius can be interpolated within a prescribed interval. In particular, mining *stable* features of geometric domains in this one-parameter family hints at important bio-physical structures. This work puts us in position to make a precise assessment of putatives models of the Nuclear Pore Complex (NPC) [21], the largest protein assembly known to date in eucaryotic cells.

From a geometric and topological standpoint, the methodology consists of the α-shapes associated to a compoundly-weighted Voronoi diagram — whose bisectors are degree-four algebraic surfaces.

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 [45]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [48]. Current investigations follow two routes. From the experimental perspective [30], directed mutagenesis enables one to quantify the energetic

[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.

importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [42]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [37].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change [2], or a function stating that a complex is more stable than another one, in which case the value returned is a score and not an energy. Borrowing to statistical mechanics [22], the usual way to design scoring functions is to mimic the so-called potentials of mean force. To put it briefly, one reverts Boltzmann's law, that is, denoting $p_i(r)$ the probability of two atoms –defining type i– to be located at distance r, the (free) energy assigned to the pair is computed as $E_i(r) = -kT \log p_i(r)$. Estimating from the PDB one function $p_i(r)$ for each type of pair of atoms, the energy of a complex is computed as the sum of the energies of the pairs located within a distance threshold [46], [33]. To compare the energy thus obtained to a reference state, one may compute $E = \sum_i p_i \log p_i/q_i$, with p_i the observed frequencies, and q_i the frequencies stemming from an a priori model [38]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

Methodological developments. 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 [10]. 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 [23]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [47], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond —a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [29].

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 [8]. 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 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.

In the dynamic setting, one wishes to understand whether selected (hot) residues exhibit specific dynamic properties, so as to serve as anchors in a binding process [41]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

3.3. Modeling 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

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

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 [28]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

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

Given a sampling on an energy landscape, a number of fundamental issues actually arise: how does the point cloud describe the topography of the energy landscape (a question reminiscent from Morse theory)? can one infer the effective number of degrees of freedom of the system over the simulation, and is this number varying? Answers to these questions would be of major interest to refine our understanding of folding and docking, with applications to the prediction of structural properties. It should be noted in passing that such questions are probably related to modeling phase transitions in statistical physics where geometric and topological methods are being used [39].

From an algorithmic standpoint, such questions are reminiscent of *shape learning*. Given a collection of samples on an (unknown) *model*, *learning* consists of guessing the model from the samples —the result of this process may be called the *reconstruction*. In doing so, two types of guarantees are sought: topologically speaking, the reconstruction and the model should (ideally!) be isotopic; geometrically speaking, their Hausdorff distance should be small. Motivated by applications in Computer Aided Geometric Design, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years [6]. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison [24]; the study of Morse-like constructions stemming from distance functions to points [32]; the analysis of topological invariants of the model and the samples, and their comparison [25], [26].

Last but not 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.

In collaboration with S. Loriot, from the GEOMETRY FACTORY.

Modeling the interfaces of macro-molecular complexes is key to improve our understanding of the stability and specificity of such interactions. We proposed a simple parameter-free model for macro-molecular interfaces, which enables a multi-scale investigation —from the atomic scale to the whole interface scale. As discussed in [10] and [2], this interface model improves the state-of-the-art to (i) identify interface atoms, (ii) define interface patches, (iii) assess the interface curvature, (iv) investigate correlations between the interface geometry and water dynamics / conservation patterns / polarity of residues.

The corresponding software, *Intervor*, has been made available to the community from the web site http://cgal.inria.fr/abs/Intervor. This software is presented in the following application note [12] . To the best of our knowledge, this software is the only publicly available one for analyzing (Voronoi) interfaces in macromolecular complexes.

4.1.2. Computing Molecular Surfaces and Volumes

Participant: Frédéric Cazals.

In collaboration with S. Loriot, from the GEOMETRY FACTORY.

Molecular surfaces and volumes are paramount to molecular modeling, with applications to electrostatic and energy calculations, interface modeling, scoring and model evaluation, pocket and cavity detection, etc. However, for molecular models represented by collections of balls (Van der Waals and solvent accessible models), such calculations are challenging in particular regarding numerics. Because all available programs are overlooking numerical issues, which in particular prevents them from qualifying the accuracy of the results returned, we developed the first certified algorithm. The corresponding piece of code uses so-called certified predicates to guarantee the branching operations of the program, as well as interval arithmetic to return an interval certified to contain the exact value of each statistic of interest—in particular the exact surface area and the exact volume of the molecular model processed. (As of December 2010, the corresponding publication is under revision.)

The corresponding software, *Vorlume*, has been made available to the community from the web site http://cgal.inria.fr/abs/Vorlume.

4.1.3. Protein Structure Comparison by Contact Map Overlap Maximization

Participant: Noël Malod-Dognin.

In collaboration with N. Yanev, University of Sofia, and IMI at Bulgarian Academy of Sciences, Bulgaria, and R. Andonov, INRIA Rennes - Bretagne Atlantique, and IRISA/University of Rennes 1, France.

Structural similarity between proteins provides significant insights about their functions. Maximum Contact Map Overlap maximization (CMO) received sustained attention during the past decade and can be considered today as a credible protein structure measure. This paper [19] presents A_purva, an exact CMO solver that is both efficient (notably faster than the previous exact algorithms), and reliable (providing accurate upper and lower bounds of the solution). These properties make it applicable for large-scale protein comparison and classification.

The software is made available from http://apurva.genouest.org.

5. New Results

5.1. Modeling Interfaces and Contacts

5.1.1. Revisiting the Voronoi description of Protein-Protein interfaces: Algorithms

Participant: Frédéric Cazals.

Describing macro-molecular interfaces is key to improve our understanding of the specificity and of the stability of macro-molecular interactions, and also to predict complexes when little structural information is known. Ideally, an interface model should provide easy-to-compute geometric and topological parameters exhibiting a good correlation with important bio-physical quantities. It should also be parametric and amenable to comparisons. In this spirit, we developed an interface model based on Voronoi diagrams, which proved instrumental to refine state-of-the-art conclusions and provide new insights [14].

This work formally presents this Voronoi interface model. First, we discuss its connexion to classical interface models based on distance cut-offs and solvent accessibility. Second, we develop the geometric and topological constructions underlying the Voronoi interface, and design efficient algorithms based on the Delaunay triangulation and the α-complex.

We conclude with perspectives. In particular, we expect the Voronoi interface model to be particularly well suited for the problem of comparing interfaces in the context of large-scale structural studies.

5.2. Algorithmic Foundations

5.2.1. Multi-scale Geometric Modeling of Ambiguous Shapes with Toleranced Balls and Compoundly Weighted α-shapes

Participants: Frédéric Cazals, Tom Dreyfus.

Dealing with ambiguous data is a challenge in Science in general and geometry processing in particular. One route of choice to extract information from such data consists of replacing the ambiguous input by a continuum, typically a one-parameter family, so as to mine stable geometric and topological features within this family. This work follows this spirit and introduces a novel framework to handle 3D ambiguous geometric data which are naturally modeled by balls [15].

First, we introduce *toleranced balls* to model ambiguous geometric objects. A toleranced ball consists of two concentric balls, and interpolating between their radii provides a way to explore a range of possible geometries. We propose to model an ambiguous shape by a collection of toleranced balls, and show that the aforementioned radius interpolation is tantamount to the growth process associated with an additively-multiplicatively weighted Voronoi diagram (also called compoundly weighted or CW). Second and third, we investigate properties of the CW diagram and the associated CW α-complex, which provides a filtration called the λ-complex. Fourth, we sketch a naive algorithm to compute the CW VD. Finally, we use the λ-complex to assess the quality of models of large protein assemblies, as these models inherently feature ambiguities.

6. Partnerships and Cooperations

6.1. European Initiatives

6.1.1. ICT FET open project: Computational Geometric Learning (CGL)

Participants: Frédéric Cazals, Christine-Andrea Roth.

This CGL project, see http://cglearning.eu/ and http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_LANG=IT&PJ_RCN=11500662, will be operated from 11/01/2010 to 10/31/2013. It is run in collaboration with Jena Univ. (coord.), INRIA (Geometrica Sophia, Geometrica Saclay, ABS), Tech. Univ. of Dortmund, Tel Aviv Univ., Nat. Univ. of Athens, Univ. of Groningen, ETH Zürich, Freie Univ. Berlin.

High dimensional geometric data are ubiquitous in science and engineering, and thus processing and analyzing them is a core task in these disciplines. The Computational Geometric Learning project (CG Learning or CGL) aims at extending the success story of geometric algorithms with guarantees, as achieved in the CGAL library and the related EU funded research projects, to spaces of high dimensions. This is not a straightforward task. For many problems, no efficient algorithms exist that compute the exact solution in high dimensions. This behavior is commonly called the curse of dimensionality.

We plan to address the curse of dimensionality by focusing on inherent structure in the data like sparsity or low intrinsic dimension, and by resorting to fast approximation algorithms. The following two kinds of approximation guarantee are particularly desirable: first, the solution approximates an objective better if more time and memory resources are employed (algorithmic guarantee), and second, the approximation gets better when the data become more dense and/or more accurate (learning theoretic guarantee). To lay the foundation of a new field—computational geometric learning—we will follow an approach integrating both theoretical and practical developments, the latter in the form of the construction of a high quality software library and application software.

In this context, the contribution of ABS lies in the work-package *Modeling high-dimensional geometric structures in science and engineering*, and is concerned with the investigation of so-called *energy landscapes*, which are hyper-surfaces describing the energetic behavior of macro-molecular systems as a function of conformational variables.

7. Dissemination

7.1. Animation of the Scientific Community

7.1.1. Conference Program Committees

– F.Cazals was member of the following PC:

- Symposium on Geometry Processing
- International conference on Pattern Recognition in Bioinformatics

7.1.2. Ph.D. thesis and HDR Committees

– F.Cazals acted as *rapporteur* for the following PhD thesis defenses:

- Duc Thanh Le, University of Toulouse, October 2010, *Rapporteur*. Thesis subject: *(Dis)assembly path planning for complex objects and applications to structural biology*. Advisors: T. Siméon and J. Cortès.
- Noël Malod-Dognin, Univ. of Rennes, January 2010, *Rapporteur*. Thesis subject: *Protein structure comparison : From Contact Map Overlap Maximisation to Distance-Based Alignment Search Tool* Advisor: R. Andonov.
- Mathias Carlen, EPFL, January 2010, *Rapporteur*. Thesis subject: *Computation and Visualization of Ideal Knot Shapes* Advisor: J. Maddocks.

7.1.3. Appointments

– F. Cazals has been appointed in the scientific committe of *GDR Bio-informatique-Moléculaire*, in charge of activities related to computational structural biology.

7.2. Teaching

7.2.1. Teaching Responsibilities

F. Cazals is co-coordinator or the *Master of Science in Computational Biology*, University of Nice - Sophia-Antipolis. This master provides an advanced curriculum at the interface of biology, computer science and applied mathematics, and is geared towards an international audience. See http://www.computationalbiology.eu.

7.2.2. Teaching at Universities

– Ecole Centrale Paris, 3rd year (master); Introduction to Computational Structural Biology; F. Cazals, 24h.
– University of Nice - Sophia-Antipolis, Master of Science in Computational Biology; Algorithmic Problems in Computational Structural Biology; F. Cazals, 24h.

7.2.3. Internships

Internship proposals can be seen on the web from the Positions section at http://www-sop.inria.fr/abs/
– Christine-Andrea Roth, MVA Cachan; Master internship: *Designing collective coordinates*; Advisor: F. Cazals; Co-advisor: C. Robert, IBPC Paris.
– Muhammad Ammad Ud Din; MSc Computational Biology, Univ. of Nice; Master internship: *Modeling antibody / antigen binding patches*; Advisor: F. Cazals.
– Achin Bansal, IIT Bombay; Summer internship: *Modeling protein binding patches*; Advisor: F. Cazals.

– Palak Dalal, IIT Bombay; Summer internship: *Geometric optimization problems for collections of balls*; Advisor: F. Cazals.

7.2.4. Ongoing Ph.D. theses

– Tom Dreyfus, university of Nice Sophia-Antipolis; Topic: *Modeling large macro-molecular assemblies*; Advisor: F. Cazals.
– Christine-Andrea Roth, university of Nice Sophia-Antipolis; Topic: *Revisiting macro-molecular flexibility, with applications to docking*; Advisor: F. Cazals.

7.3. Participation to Conferences, Seminars, Invitations

7.3.1. Invited Talks

F. Cazals gave the following invited talks:

- *Balls, sticks, triangles and molecules*, closing workshop of the ANR Triangles, Sophia-Antipolis, December 2010.
- *Assessing the stability of protein complexes within large assemblies*, EMBO Symposium on Molecular Perspectives on Protein-Protein Interactions, Sant Feliu de Guixols, Spain, November 2010.
- *Assessing the stability of protein complexes within large assemblies*, XXIIeme Congrès de la Société Française de Biophysique, La Colle sur Loup, September 2010.
- *Assessing the stability of protein complexes within large assemblies*, LAAS, Toulouse, September 2010.
- *Geometric Models for the Description of 3D Molecular Systems*, Energy Landscapes Workshop, Chemnitz, June 2010.
- *Geometric Models for the Description of High-dimensional Point Cloud Data*, Energy Landscapes Workshop, Chemnitz, June 2010.
- *Modeling the interface of protein - protein complexes: shelling the Voronoi interface reveals patterns of composition, residue conservation, and water dynamics,* IBMC Strasbourg, Architecture et réactivité de l'ARN, May 2010.
- *Modeling water traffic at protein interfaces: from Voronoi models to (simple) percolation on lattices*, Ecole Normale Supérieure, groupe de travail Probabilités et Statistiques, April 2010.

7.3.2. The ABS Seminar

The ABS seminar featured presentations from the following visiting scientists:
– Charles Robert, Institut de Biologie Physico-Chimique, Paris.
– Annick Dejaegere, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France.
– Patrick Schultz, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France.
– Sameer Velankar, European Biological Institute, UK.
– Erik Aurell, Aalto University, Helsinki, Finland, and KTH Royal Institute of Technology, Stockholm, Sweden.

8. Bibliography

Major publications by the team in recent years

[1] 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.

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

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

[4] F. CAZALS, F. CHAZAL, T. LEWINER. *Molecular shape analysis based upon the Morse-Smale complex and the Connolly function*, in "ACM SoCG", San Diego, USA, 2003.

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

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

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

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

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

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

Publications of the year

Articles in International Peer-Reviewed Journal

[11] S. LORIOT, F. CAZALS, J. BERNAUER. *ESBTL: efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules*, in "Bioinformatics", 2010, vol. 26, n^{o} 8, p. 1127–1128.

[12] S. LORIOT, F. CAZALS. *Modeling Macro-Molecular Interfaces with Intervor*, in "Bioinformatics", 2010, vol. 26, n^{o} 7, p. 964–965.

[13] N. MALOD-DOGNIN, R. ANDONOV, N. YANEV. *Solving Maximum Clique Problem for Protein Structure Similarity*, in "Serdica Journal of Computing", 2010, vol. 4, n^{o} 1, p. 93–100.

International Peer-Reviewed Conference/Proceedings

[14] F. CAZALS. *Revisiting the Voronoi description of Protein-Protein interfaces: Algorithms*, in "IPAR International Conference on Pattern Recognition in Bioinformatics", Nijmegen, the Netherlands, T. DIJKSTRA, E.

TSIVTSIVADZE, E. MARCHIORI, T. HESKES (editors), Lecture Notes in Bioinformatics #6282, 2010, p. 419–430.

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

[16] N. MALOD-DOGNIN, R. ANDONOV, N. YANEV. *Maximum Clique in Protein Structure Comparison*, in "9th International Symposium on Experimental Algorithms", Ischia Island, Italy, P. FESTA (editor), Springer Berlin / Heidelberg, 2010, p. 106–117.

[17] L. MAVRIDIS, V. VENKATRAMAN, D. W. RITCHIE, N. MORIKAWA, R. ANDONOV, A. CORNU, N. MALOD-DOGNIN, J. NICOLAS, M. TEMERINAC-OTT, M. REISERT, H. BURKHARDT, A. AXENOPOULOS. *SHREC-10 Track: Protein Models*, in "3DOR: Eurographics Workshop on 3D Object Retrieval", Norrkoping, Sweden, I. PRATIKAKIS, M. SPAGNUOLO, T. THEOHARIS, R. VELTKAMP (editors), The Eurographics Association, 2010, p. 117–124.

Research Reports

[18] F. CAZALS, T. DREYFUS. *Multi-scale Geometric Modeling of Ambiguous Shapes with Toleranced Balls and Compoundly Weighted alpha-shapes*, INRIA, Jul 2010, RR-7306, http://hal.inria.fr/inria-00497688, http://.

[19] N. MALOD-DOGNIN, N. YANEV, R. ANDONOV. *Comparing Protein 3D Structures Using A_purva*, INRIA, 11 2010, n^{o} RR-7464, http://hal.inria.fr/inria-00539939/PDF/RR-7464.pdf.

References in notes

[20] 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, p. 683-694.

[21] 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, p. 695–701.

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

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

[24] F. CHAZAL, D. COHEN-STEINER, A. LIEUTIER. *A sampling theory for compact sets in Euclidean space*, in "Discrete and Computational Geometry", 2009, vol. 41, n^{o} 3, p. 461–479.

[25] F. CHAZAL, A. LIEUTIER. *Weak Feature Size and persistent homology : computing homology of solids in $\mathbb{R}^n$ from noisy data samples*, in "ACM SoCG", 2005, p. 255-262.

[26] D. COHEN-STEINER, H. EDELSBRUNNER, J. HARER. *Stability of Persistence Diagrams*, in "ACM SoCG", 2005.

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

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

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

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

[31] 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, p. 531–539.

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

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

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

[35] V. K. KRIVOV, M. KARPLUS. *Hidden complexity of free energy surfaces for peptide (protein) folding*, in "PNAS", 2004, vol. 101, n^{o} 41, p. 14766-14770.

[36] E. MEERBACH, C. SCHUTTE, I. HORENKO, B. SCHMIDT. *Metastable Conformational Structure and Dynamics: Peptides between Gas Phase and Aqueous Solution*, in "Analysis and Control of Ultrafast Photoinduced Reactions. Series in Chemical Physics 87", O. KUHN, L. WUDSTE (editors), Springer, 2007.

[37] I. MIHALEK, O. LICHTARGE. *On Itinerant Water Molecules and Detectability of Protein-Protein Interfaces through Comparative Analysis of Homologues*, in "JMB", 2007, vol. 369, n^{o} 2, p. 584–595.

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

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

[40] E. PLAKU, H. STAMATI, C. CLEMENTI, L. KAVRAKI. *Fast and Reliable Analysis of Molecular Motion Using Proximity Relations and Dimensionality Reduction*, in "Proteins: Structure, Function, and Bioinformatics", 2007, vol. 67, n^{o} 4, p. 897–907.

[41] D. RAJAMANI, S. THIEL, S. VAJDA, C. CAMACHO. *Anchor residues in protein-protein interactions*, in "PNAS", 2004, vol. 101, n^{o} 31, p. 11287-11292.

[42] D. REICHMANN, O. RAHAT, S. ALBECK, R. MEGED, O. DYM, G. SCHREIBER. *From The Cover: The modular architecture of protein-protein binding interfaces*, in "PNAS", 2005, vol. 102, n^{o} 1, p. 57-62 [*DOI :* 10.1073/PNAS.0407280102], http://www.pnas.org/cgi/content/abstract/102/1/57.

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

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

[45] G. SCHREIBER, L. SERRANO. *Folding and binding: an extended family business*, in "Current Opinion in Structural Biology", 2005, vol. 15, n^{o} 1, p. 1–3.

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

[47] C. SUMMA, M. LEVITT, W. DEGRADO. *An atomic environment potential for use in protein structure prediction*, in "JMB", 2005, vol. 352, n^{o} 4, p. 986–1001.

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