INRIA

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 [Team leader; DR2 Inria, HdR]
Julie Bernauer [CR2 Inria; Visiting the AMIB INRIA team from September 1st, 2009.]

PhD Student

Tom Dreyfus [MESR monitor fellow]

Post-Doctoral Fellow

Sébastien Loriot [INRIA - Direction Scientifique]

Administrative Assistant

Caroline French [*Fonctionnaire stagiaire*, assistant of GEOMETRICA and ABS]

Other

Ludovic Di Benedetto [Master student, Univ. of Montpellier, March - August 2009]
Harshad Kanhere [Summer intern from IIT Bombay - India, May-July 2009]
Emilie Pihan [Master student, Univ. of Nice-Sophia-Antipolis, March-August 2009]
Nisarg Shah [Summer intern from IIT Bombay - India, May-July 2009]

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 [47]. 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 [45], [32] and later Connolly [28], while defining molecular surfaces and developing algorithms to compute them, established the connexions between molecular modeling and geometric constructions. Remarkably, a number of difficult problems (e.g. additively weighted Voronoi diagrams) were touched upon in these early days.

The models developed in this vein are instrumental in investigating the interactions of molecules for which no structural data is available. But such models often fall short from providing complete answers, which we illustrate with the folding problem. On one hand, as the conformations of side-chains belong to discrete sets (the so-called rotamers or rotational isomers) [36], 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.

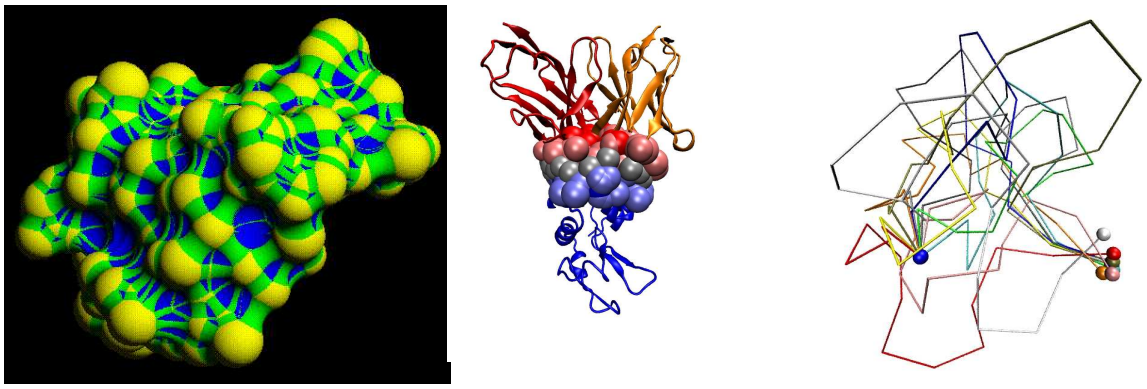

Figure 1. (a) Molecular surface (b) An antibody-antigen complex, with interface atoms computed as described in [10] (c)Conformations of a backbone loop

2.2. Highlights of the year

Because all biological functions are accounted for by macro-molecular complexes, modeling the interfaces between the partners found within this complexes is a key endeavour. To carry out this modeling, fundamental

biophysical parameters are the geometry and the structure of the interfaces [44], their chemical composition in terms of amino-acids [22], the dynamics of the water molecules in-between the partners [39], as well as the conservation of residues [35].

We introduced a geometric model for interfaces based on the α-complex [10], and proposed a process which amounts to shelling this model into concentric shells [11]. Doing so enables the definition of a notion of *depth* of atoms at the interface, from which the study of correlations between the aforementioned quantities can be significantly refined. The paper made the front cover of the journal *Proteins*, which is a reference venue for structural biology. See Fig. 2.

While our results are substantiated by statistical calculations (p-values), they actually call for further development in the realm of percolation theory, so as to provide quantitative physical models describing the behavior of the solvent molecules at the interface. This bridges the gap between an important problem in biophysics, and state-of-the-art developments in mathematics—cf the work of the Fields medalist Wendelin Werner.

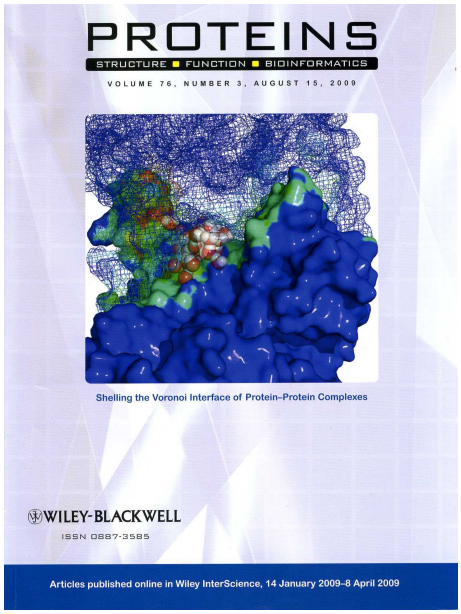

Figure 2. Modeling the bahaviour of water molecules at protein - protein interfaces. Cover of the Proteins journal. See [11].

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 [47]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [50]. Current investigations follow two routes. From the experimental perspective [31], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [44]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [39].

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 [23], 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 [48], [34]. 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 [40]. 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 [24]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [49], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond —a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [30].

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 [13]. 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.

[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 dynamic setting, one wishes to understand whether selected (hot) residues exhibit specific dynamic properties, so as to serve as anchors in a binding process [43]. 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 [29]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [46]. 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 [42], to Morse theory [37] and to analysis of meta-stable states of time series [38] 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 such questions are probably related to modeling phase transitions in statistical physics where geometric and topological methods are being used [41].

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 [7]. 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 [25]; the study of Morse-like constructions stemming from distance functions to points [33]; the analysis of topological invariants of the model and the samples, and their comparison [26], [27].

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

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.

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 [11], 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 [21]. To the best of our knowledge, this code is the only publicly available one for analyzing (Voronoi) interfaces in macro-molecular complexes.

4.1.2. Discrimination between crystallographic and biological protein-protein interactions

Participant: Julie Bernauer.

Knowing the oligomeric state of a protein is necessary to understand its function. This tool, accessible as a webserver http://cgal.inria.fr/DiMoVo, provides a reliable discrimination function to obtain the most favorable state of proteins. See [2].

4.1.3. Protein-protein docking conformation evaluation

Participant: Julie Bernauer.

Scoring is a crucial part of a protein-protein procedure and having a quantitative function to evaluate conformations is mandatory. This server http://cgal.inria.fr/VorScore provides access to a geometric knowledge-based evaluation function. See [1] for further details.

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. Assessing the Stability of Protein Complexes within Large Assemblies

Keywords: *Tandem Affinity Purification*, *curved Voronoi diagrams*, *data integration*, *large protein assemblies*, *topological stability*.

Participants: Frédéric Cazals, Tom Dreyfus.

Structural genomics projects, in particular those exploiting Tandem Affinity Purification (TAP), have revealed remarkable features of full proteomes. While these insights are essentially of combinatorial nature, that is a number of proteins are known to interact within a complex, leveraging this information will require building three dimensional models of these assemblies. Such an endeavour has recently been completed for the Nuclear Pore Complex, for which plausible reconstructions have been computed from different experimental data, including TAP data. But the reconstruction is qualitative and the coherence to TAP data is not analyzed.

In this work [16], we introduce toleranced collections of balls to represent protein assemblies known with uncertainties, together with a method highlighting stable complexes within such assemblies. The method relies on the computation of the topological stability of the connected components in a collection of balls growing according to a so-called additively-multiplicatively-weighted Voronoi diagram. In particular, our strategy enables the investigation of the coherence between a reconstructed model and TAP data.

5.1.2. Comparing Voronoi and Laguerre tessellations in the protein-protein docking context

Keywords: *Protein-protein interactions*, *Voronoi and Laguerre tessellations*, *docking*, *learning*.

Participant: Julie Bernauer.

In collaboration with Thomas Bourquard, Jerome Azé, and Anne Poupon.
T. Bourquard is with LRI Université Paris-Sud 11.
J. Azé is with LRI Université Paris-Sud 11.
A. Poupon is in the Physiologie de la Reproduction et des Comportements lab, INRA Tours.

Most proteins fulfill their functions through the interaction with other proteins. Because most of these interactions are transitory, they are difficult to detect experimentally, and obtaining the structure of the complex is generally not possible. Consequently, prediction of the existence of these interactions and of the structure of the resulting complex has received a lot of attention in the last decade. However, proteins are very complex objects, and classical computing methods have led to computer-time consuming methods, whose accuracy is not sufficient for large scale exploration of the so-called “interactome”, the ensemble of protein-protein complexes in the cell. In order to design an accurate and high-throughput prediction method for protein-protein docking, the first step was to model a protein structure using a formalism amenable to fast computation, without losing the intrinsic properties of the object. In our work [14], we have tested two different, but related, formalisms: the Voronoi and Laguerre tessellations. We present here a comparison of these two models in the context of protein-protein docking.

5.1.3. A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation

Keywords: *Knowledge-based potential*, *coarse-grained model*, *spherical arrangements*, *structure prediction and refinement*, *surface area*.

Participants: Julie Bernauer, Frédéric Cazals, Sébastien Loriot.

In collaboration with M. Levitt, Dpt of Structural Biology, Stanford University.

Knowledge-based protein folding potentials have proven successful in the recent years. Based on statistics of observed interatomic distances, they generally encode pairwise contact information.

In this study [15], we present a method that derives multi-body contact potentials from measurements of surface areas using coarse-grained protein models. The measurements are made using a newly implemented geometric construction: the arrangement of circles on a sphere [13]. This construction enables the definition of residue covering areas which are used as parameters to build functions able to distinguish native structures from decoys. These functions, encoding up to 5-body contacts are evaluated on a reference set of 66 structures and its 45000 decoys, and also on the often used *lattice_ssfit* set from the *decoys'R* us database. We show that the most relevant information for discrimination resides in 2- and 3-body contacts. The potentials we have obtained can be used for evaluation of putative structural models; they could also lead to different types of structure refinement techniques that use multi-body interactions.

5.2. Modeling the flexibility of macro-molecules

As a follow-up to [17], we started preliminary work on the problem of dimensionality reduction for energy landscapes.

5.3. Algorithmic foundations

5.3.1. Computing the Volume of a Union of Balls: a Certified Algorithm

Participants: Frédéric Cazals, Sébastien Loriot.

In collaboration with H. Kanhere, master student at the Indian Institute of Technology, Bombay.

Balls and spheres are amongst the simplest $3D$ modeling primitives, and computing the volume of a union of balls is an elementary problem. Although a number of strategies addressing this problem have been investigated in several communities, we are not aware of any robust algorithm, and present the first such algorithm [20].

Our calculation relies on the decomposition of the volume of the union into convex regions, namely the restrictions of the balls to their regions in the power diagram. Theoretically, we establish a formula for the volume of a restriction, based on Gauss' divergence theorem. The proof being constructive, we develop the associated algorithm. On the implementation side, we carefully analyse the predicates and constructions involved in the volume calculation, and present a certified implementation relying on interval arithmetic. The result is certified in the sense that the exact volume belongs to the interval computed using the interval arithmetic.

Experimental results are presented on hand-crafted models presenting various difficulties, as well as on the 58,898 models found in the 2009-07-10 release of the Protein Data Bank.

5.3.2. Reconstructing 3D compact sets

Participant: Frédéric Cazals.

In collaboration with D. Cohen-Steiner, from Geometrica, INRIA Sophia-Antipolis.

Reconstructing a 3D shape from sample points is a central problem faced in medical applications, reverse engineering, natural sciences, cultural heritage projects, etc. While these applications motivated intense research on 3D surface reconstruction, the problem of reconstructing more general shapes hardly received any attention. This paper [19] develops a reconstruction algorithm changing the 3D reconstruction paradigm as follows.

First, the algorithm handles general shapes i.e. compact sets as opposed to surfaces. Under mild assumptions on the sampling of the compact set, the reconstruction is proved to be correct in terms of homotopy type. Second, the algorithm does not output a single reconstruction but a nested sequence of *plausible* reconstructions. Third, the algorithm accommodates topological persistence so as to select the most stable features only. Finally, in case of reconstruction failure, it enables the identification of under-sampled areas, so as to possibly fix the sampling.

These key features are illustrated by experimental results on challenging datasets, and should prove instrumental in enhancing the processing of such datasets in the aforementioned applications.

5.3.3. Robust and Efficient Delaunay triangulations of points on or close to a sphere

Participant: Sébastien Loriot.

In collaboration with Manuel Caroli, Pedro M.M. de Castro, Monique Teillaud, and Camille Wormser. M. Caroli, P. M.M. de Castro and M. Teillaud are with Geometrica, INRIA Sophia-Antipolis. C. Wormser is with the CS Dpt, ETH Zurich.

We propose two approaches for computing the Delaunay triangulation of points on a sphere, or of rounded points close to a sphere [18]. Both approaches are based on the classic incremental algorithm initially designed for the plane. The space of circles gives the mathematical background for this work. We implemented the two approaches in a fully robust way, building upon existing generic algorithms provided by the cgal library. The effciency and scalability of the method is shown by benchmarks

5.4. Software

5.4.1. Investigating Interfaces of Macro-Molecular Complexes with Intervor

Participants: Frédéric Cazals, Sébastien Loriot.

`Intervor` is a software computing a parameter free representation of macro-molecular interfaces, based on the α-complex of the atoms [21]. Given two interacting partners, possibly with water molecules squeezed in-between, `Intervor` computes an interface model which has the following characteristics: (i) it identifies the atoms of the partners which are in direct contact and those whose interaction is water mediated, (ii) it defines a geometric complex separating the partners, the Voronoi interface, whose geometric and topological descriptions are straightforward (surface area, number of patches, curvature), (iii) it enables the definition of the depth of atoms at the interface, thus going beyond the traditional dissection of an interface into a core and a rim.

These features can be used to investigate correlations between structural parameters and key properties such as the conservation of residues, their polarity, the water dynamics at the interface, mutagenesis data, etc.

`Intervor` can be run from the web site http://cgal.inria.fr/abs/Intervor, or in stand-alone mode upon downloading the binary file. Plugins are also made available for Visual Molecular Dynamics (VMD) and Pymol.

5.4.2. ESBTL: Easy Structural Biology Template Library

Participants: Julie Bernauer, Frédéric Cazals, Sébastien Loriot.

The ever increasing number of structural biological data calls for robust and efficient software for analysis. ESBTL (Easy Structural Biology Template Library) is a lightweight C++ library that handles PDB data and provides a data structure suitable for geometric analysis and advanced constructions. The parser and data model provided by this ready-to-use header-only library provide adequate treatment of usually discarded information (insertion codes, atom occupancy...) while still being able to detect badly formatted files. The template-based structure enables rapid design of new computational structural biology applications and is fully compatible with the new remediated PDB archive format. It also makes the code easy-to-use while being versatile enough to allow advanced user developments.

ESBTL is freely available under the GNU General Public License from http://esbtl.sf.net. The website provides the source code, examples, code snippets and documentation.

6. Other Grants and Activities

6.1. International initiatives

6.1.1. Associated team GNAPI

Participants: Julie Bernauer, Frédéric Cazals, Sébastien Loriot.

This project is a collaboration between members of the ABS group in Sophia-Antipolis and the Levitt group in Stanford University on geometric and knowledge-based analysis for large biomolecules and their interactions. The function of a biomolecule largely depends on its interaction with other partners such as proteins and nucleic acids. To better understand and to be able to predict the behavior of these macromolecules we offer an innovative approach combining physico-chemical properties description, distance measurements, geometrical descriptors and a statistical description of the system. This work is based on the recent developments on knowledge-based potentials and Voronoi/alpha-shape protein descriptions. They offer nice tools to get valuable insight on how the macromolecules interact which could lead to interesting therapeutics applications. Both groups involved have complementary experience in these domains. In 2009, the emphasis has been made on sphere arrangements as discriminative parameters. Results obtained were partly published in the JOBIM Conference Proceedings (See http://www.jobim2009.fr) and a presentation was made by Julie Bernauer at the conference in June 2009 in Nantes (see: [15]). Results were also presented as a poster by Sébastien Loriot at the *Flexibility and Biological Recognition: from Biophysics to Data Models* in Sophia Antipolis in March 2009 and by Julie Bernauer as a poster in ISMB/ECCB Conference in Stockholm in June 2009. Other topics are described on the website: http://www-sop.inria.fr/members/Julie.Bernauer/EquipeAssociee/Formulaire2009.htm and the form for 2010, including the work done in 2009 is available at: https://www-sop.inria.fr/members/Julie.Bernauer/EquipeAssociee/DR:I/Prolong2010.htm

7. Dissemination

7.1. Animation of the scientific community

7.1.1. Conference program committees

J.Bernauer was member of the ISMB/ECCB 09 Paper Committee (Protein Structure).

7.1.2. Ph.D. thesis and HDR committees

Frédéric Cazals was reviewer for the Ph.D. thesis of Adrien Saladin, Univ. of Paris VII, and for that of Benjamin Schwarz, Univ. of Strasbourg.

7.1.3. Conference organization

In May 2009, J. Bernauer and F. Cazals organized the conference *Flexibility and Biological Recognition: from Biophysics to Data Models*, within the scope of the EU Coordinated Action FocusK3D. The event was centered on the question of modeling the flexibility of macro-molecules, and consisted of 18 invited talks. See http://www-sop.inria.fr/manifestations/fmr2009/program_en.shtml for a description of the event.

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

– École Polytech'Nice, Engineering curriculum (3rd year), Introduction à la "Biogéométrie": J.Bernauer (3h).

– Université de Nice Sophia Antipolis / University of Bologna Winter School, Algorithms and structural biology: Introduction to biogeometry and other computational structural biology techniques: J.Bernauer (3h).

– AgroParisTech, Paris, MAP3 (module d'approfondissement) Ingénierie des protéines, cursus ingénieur agronome, deuxième année; Introduction à la bioinformatique; J.Bernauer (3h).

– Master Bioinformatique et Biostatistiques (BIBS), Orsay University; Algorithmic Problems in Computational Structural Biology; F. Cazals (12h), J. Janin (6h), C. Robert (6h).
– University of Nice - Sophia-Antipolis, Master of Science in Computational Biology; Algorithmic Problems in Computational Structural Biology; J. Bernauer (9h), F. Cazals (15h).

7.2.3. Internships

Internship proposals can be seen on the web from the Positions section at http://www-sop.inria.fr/abs/
– Ludovic Di Benedetto; *Development of scoring functions for protein-nucleic acid docking*; Univ. of Montpellier 2, Master 2 in Biostatistics. Advisor: J. Bernauer.

– Harshad Kanhere; *Computing the Volume of Union of Balls*, IIT Bombay. Advisor: F. Cazals.

– Emilie Pihan; *Study of surface methods to identify similarities between ligands*; École Polytechnique Nice - Univ. of Nice - Sophia-Antipolis. Advisor: J. Bernauer and D. Douguet (CNRS / IPMC Sophia Antipolis).

– Nisarg Shah; *Geometric optimization problems for collections of balls*; IIT Bombay. Advisor: F. Cazals.

7.2.4. Ongoing Ph.D. theses

– Tom Dreyfus, *Modeling large macro-molecular assemblies*, university of Nice Sophia-Antipolis.

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.

– *A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation*; J.Bernauer, INRIA-NIH meeting, June 2009.

– *Computational Structural Biology:Periodic Triangulations for Molecular Dynamics*; J.Bernauer, Workshop "Subdivide and tile: Triangulating spaces for understanding the world", organized in Leiden (Netherlands), 16-20 November, 2009. See: http://www.lorentzcenter.nl/lc/web/2009/357/info.php3?wsid=357.

– *Geometric techniques for the inference and the assessment of macro-molecular complexes*; F. Cazals, IGBMC Strasbourg, September 2009.

– *Geometric representations of protein complexes and assemblies: an excursion across scales*; F. Cazals, INRIA - NIH meeting, June 2009.

– *Models and algorithms for the description of macro-molecular interactions*; F. Cazals Workshop Industry Challenges in Geometric Modeling, CAD, and Simulation; Darmsdadt, March 2009.

7.3.2. The ABS seminar

The following people gave invited talks within the scope of the workshop *Flexibility in biological recognition* organized by J. Bernauer and F. Cazals:
– A. Carbone, Univ. Pierre et Marie Curie, Paris, France
– P. Chakrabarti, Bose Institute, Kolkata, India
– O. Corby, INRIA, Sophia Antipolis, France
– J. Cortes, LAAS, Toulouse, France
– S. Fiorucci, Univ. of Nice, France
– J. Janin, Univ. Paris-Sud, Orsay, France
– R. Lavery, IBCP, Lyon, France
– M. Levitt, Stanford University, USA
– E. Lindahl, Stockholm University, Sweden
– A. Poupon, INRA, Tours, France
– C. Prevost, IBPC, Paris, France
– S. Redon, INRIA, Grenoble, France
– C. Robert, Univ. Paris-Sud, Orsay, France
– M. Spagnuolo, CNR/IMATI, Genova, Italy

The ABS seminar featured presentations from the following visiting scientists:
– Olivier Gibaru, Arts & Métiers ParisTech – Lille, France
– Olivier Lichtarge, Baylor College of Medicine, Texas, USA
– Natasa Przulj, Imperial College London, Dpt of Computing, UK

7.3.3. Scientific visits

ABS has hosted the following scientists:
– For the Équipe Associée GNAPI collaboration, Michael Levitt visited ABS in Sophia in March for the FMR2009 Workshop. Adelene Sim, PhD candidate at Stanford University, visited J.Bernauer in the AMIB team for two weeks in October. She is working in collaboration on RNA folding knowledge-based force-fields.

7.3.4. Misc activities

– J. Bernauer was appointed member of the CUMIR in INRIA Sophia Antipolis for 2009. She had to resign as she was moving on September 1st, 2009.

– J. Bernauer is member of the "Groupe de Travail Stockage" at INRIA Sophia Antipolis in 2009.

8. Bibliography

Major publications by the team in recent years

[1] J. BERNAUER, J. AZÉ, 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, R. 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.

[3] J. BERNAUER, A. POUPON, J. AZÉ, 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.

[4] 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.

[5] 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.

[6] 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, San Diego, USA", 2003.

[7] 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.

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

[9] F. CAZALS, M. POUGET. *Estimating Differential Quantities using Polynomial fitting of Osculating Jets*, in "Computer Aided Geometric Design", vol. 22, n^{o} 2, 2005, 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", vol. 15, n^{o} 9, 2006, p. 2082–2092.

Year Publications

Articles in International Peer-Reviewed Journal

[11] 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", vol. 76, n^{o} 3, 2009, p. 677–692.

[12] 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", vol. 42, n^{o} 6-7, 2009, p. 536–550.

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

International Peer-Reviewed Conference/Proceedings

[14] T. BOURQUARD, J. BERNAUER, J. AZÉ, A. POUPON. *Comparing Voronoi and Laguerre tessellations in the protein-protein docking context*, in "Sixth annual International Symposium on Voronoi Diagrams, Danemark Copenhagen", F. Anton and J. Andreas Baerentzen - Technical University of Denmark, 2009-06-23, http://hal.inria.fr/inria-00429618/en/.

National Peer-Reviewed Conference/Proceedings

[15] S. LORIOT, F. CAZALS, M. LEVITT, J. BERNAUER. *A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation*, in "Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), France Nantes", Société Française de Bioinformatique, 2009-07, http://hal.inria.fr/inria-00429607/en/FR.

Workshops without Proceedings

[16] F. CAZALS, T. DREYFUS. *Assessing the Stability of Protein Complexes within Large Assemblies*, in "ISMB Satellite Meeting on Structural Bioinformatics and Computational Biophysics", 2009.

Scientific Books (or Scientific Book chapters)

[17] 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), The Institute of Mathematics and its Applications, 2009.

Research Reports

[18] M. CAROLI, P. DE CASTRO, S. LORIOT, M. TEILLAUD, C. WORMSER. *Robust and Efficient Delaunay triangulations of points on or close to a sphere*, n^{o} 7004, INRIA, 2009, http://hal.archives-ouvertes.fr/inria-00405478/en/, Technical report.

[19] F. CAZALS, D. COHEN-STEINER. *Reconstructing 3D compact sets*, n^{o} RR-6868, INRIA, 2009, http://hal.inria.fr/inria-00370208/en/, Research Report.

[20] F. CAZALS, H. KANHERE, S. LORIOT. *Computing the Volume of Union of Balls: a Certified Algorithm*, n^{o} 7013, INRIA, 2009, Technical report.

[21] F. CAZALS, S. LORIOT. *Modeling Macro-Molecular Interfaces with Intervor*, n^{o} 7069, INRIA, 2009, Technical report.

References in notes

[22] R. BAHADUR, P. CHAKRABARTI, F. RODIER, J. JANIN. *A dissection of specific and non-specific protein-protein interfaces*, in "J. Mol. Biol.", vol. 336, 2004.

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

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

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

[26] 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.

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

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

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

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

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

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

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

[34] 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.

[35] M. GUHAROY, P. CHAKRABARTI. *Conservation and relative importance of residues across protein-protein interfaces.*, in "Proc Natl Acad Sci U S A", vol. 102, n^{o} 43, Oct 2005, p. 15447–15452, http://dx.doi.org/10.1073/pnas.0505425102.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

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

[44] 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.

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

[46] 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.

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

[48] 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.

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

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