Activity Report 2011

Project-Team ABS

Algorithms, Biology, Structure

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Computational Biology and Bioinformatics

Table of contents

Project-Team ABS

Keywords: Computational Structural Biology, Protein-protein Interactions, Protein Assemblies, Computational Geometry, Computational Topology.

Beginning of the Team: 01/07/2008, End of the Team: 31/07/2012.

1. Members

Research Scientist

Frédéric Cazals [Team leader; DR2 Inria, HdR]

PhD Students

Deepesh Agarwal [INRIA, from the 10/01/2011]
Tom Dreyfus [MESR monitor fellow]
Christine Roth [INRIA CORDI-S fellow]
Alix Lhéritier [INRIA CORDI-S fellow, from the 09/01/2011]

Post-Doctoral Fellow

Noël Malod-Dognin [INRIA]

Administrative Assistant

Caroline French [Assistant of ABS, and GEOMETRICA(until the 09/01/2011), and COFFEE(since the 10/01/2011).]

Others

Pararth Shah [Summer intern from IIT Bombay - India, May-July 2011]
Venkata Duvuru [Summer intern from IIT Bombay - India, May-July 2011]

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 (DNA, RNA), proteins, small molecules) and their interactions, and come from three main sources: X ray crystallography, NMR spectroscopy, cryo Electron Microscopy. Ultimately, structural data should expand our understanding of how the structure accounts for the function of macro-molecules —one of the central questions in structural biology. This goal actually subsumes two equally difficult challenges, which are *folding* —the process through which a protein adopts its $3d$ structure, and *docking* —the process through which two or several molecules assemble. Folding and docking are driven by non covalent interactions, and for complex systems, are actually inter-twined [48]. 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 [46], [34] and later Connolly [30], 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) [37], 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.

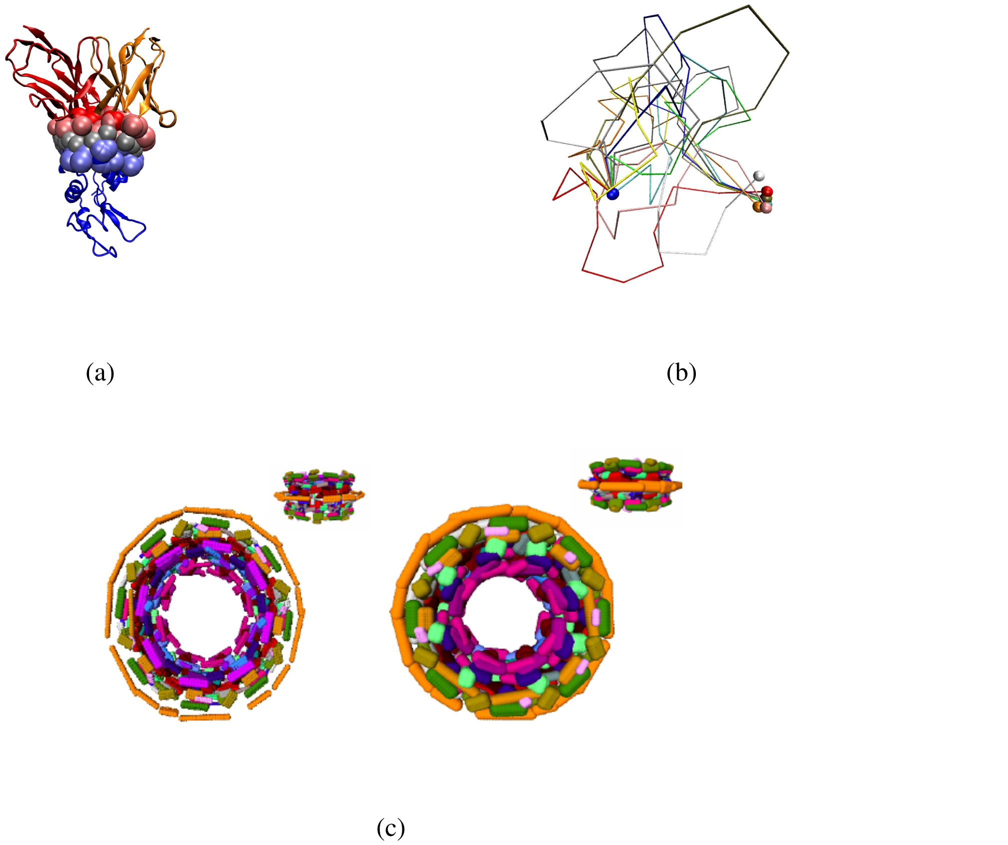
As illustrated by the previous discussion, modeling macro-molecules touches upon biology, physics and chemistry, as well as mathematics and computer science. In the following, we present the topics investigated within ABS.

2.2. Highlights of the Year

Our key achievements in 2011 have been twofold.

First, following our work on Voronoi interface models [10], [2], one of our long-standing goals has been to provide a unified model for atomic resolution protein interfaces. We took our Voronoi based modeling approach one step further, by developing a parametric model of protein binding patches, amenable to structure comparison [16], [21]. This model may be seen as a parametric *core-rim* model refining the classical binary core-rim model. It encompasses both geometric and topological properties, and allows the investigation of the topology of binding patches—a dimension ignored so far. Moreover, the topological information also makes the model amenable to structure comparison, a topic hardly touched at the atomic level—the problem is in fact NP-hard. This model is currently being used to perform a detailed analysis of antibody - antigen complexes, in the perspective of understanding the relationship between the amino-acid variability of immunoglobulins, and their binding affinity.

Second, a recent achievement has been the design of an algorithm to compute so-called compoundly-weighted Voronoi diagram, in the context of TOleranced Models [5]. Recall that the TOM framework is meant to accommodate uncertainties on the shapes and the positions of proteins within large protein assemblies. In



(a)

(b)

(c)

Figure 1. (a) An antibody-antigen complex, with interface atoms identified by our Voronoi based interface model [10], [2] (b) A diverse set of conformations of a backbone loop, selected thanks to a geometric optimization algorithm [15] (c) A Toleranced model of the nuclear pore complex, visualized at two different scales [20].

2011, we fully exploited the TOM framework to perform analysis on qualitative reconstructions of the Nuclear Pore Complex (NPC) [20], [19], the largest protein assembly known to date in the eukaryotic cell [22]. This work was carried out in collaboration with V. Doye, from Inst. Jacques Monod, Paris, a renowned expert of the NPC.

We believe that the TOM framework and the accompanying statistics should prove of general interest for the problem of reconstructing macro-molecular assemblies and that of assessing such reconstructions.

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 [48]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [51]. Current investigations follow two routes. From the experimental perspective [33], 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 [45]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [40].

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 [25], 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 [49], [36]. 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 [41]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

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

[1] For structures resolved by crystallography, the PDB contains the asymmetric unit of the crystal. Determining the biological unit from the asymmetric unit is a problem in itself.

[2] The Gibbs free energy of a system is defined by $G = H - TS$, with $H = U + PV$. G is minimum at an equilibrium, and differences in G drive chemical reactions.

In the static setting, one seeks the minimalist geometric model providing a relevant bio-physical signal. A first step in doing so consists of identifying interface atoms, so as 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 [26]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [50], 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 [32].

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 [44]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

3.3. Modeling Macro-molecular Assemblies

3.3.1. Reconstruction by data integration

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [24]. In a nutshell, the strategy is reminiscent from NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This strategy has been in particular used to propose plausible models of the Nuclear Pore Complex [23], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

3.3.2. Modeling with uncertainties and model assessment

Reconstruction by data integration requires three ingredients. First, a parametrized model must be adopted, typically a collection of balls to model a protein with pseudo-atoms. Second, as in NMR, a functional measuring the agreement between a model and the data must be chosen. In [22], this functional is based upon *restraints*, namely penalties associated to the experimental data. Third, an optimization scheme must be selected. The design of restraints is notoriously challenging, due to the ambiguous nature and/or the noise level of the data. For example, Tandem Affinity Purification (TAP) gives access to a *pullout* i.e. a list of protein types which are known to interact with one tagged protein type, but no information on the number of complexes or on the stoichiometry of proteins types within a complex is provided. In cryo-EM, the envelope enclosing an assembly is often imprecisely defined, in particular in regions of low density. For immuno-EM labelling experiments, positional uncertainties arise from the microscope resolution.

These uncertainties coupled with the complexity of the functional being optimized, which in general is non convex, have two consequences. First, it is impossible to single out a unique reconstruction, and a set of plausible reconstructions must be considered. As an example, 1000 plausible models of the NPC were reported in [22]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

3.3.3. Methodological developments

As outlined by the previous discussion, a number of methodological developments are called for. On the experimental side, the problem of fostering the interpretation of data is under scrutiny. Of particular interest is the disambiguation of proteomics signals (TAP data, mass spectrometry data), and that of density maps coming from electron microscopy. As for modeling, two classes of developments are particularly stimulating. The first one is concerned with the design of algorithms performing reconstruction by data integration. The second one encompasses assessment tools, in order to single out the reconstructions which best comply with the experimental data.

3.4. 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 [31]. 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) [47]. 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 [43], to Morse theory [38] and to analysis of meta-stable states of time series [39] 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 [42].

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

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 [27]; the study of Morse-like constructions stemming from distance functions to points [35]; the analysis of topological invariants of the model and the samples, and their comparison [28], [29].

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. Software

This section briefly comments on all the software distributed by ABS. On the one hand, the software released in 2011 is briefly described as the context is presented in the sections dedicated to new results. On the other hand, the software made available before 2011 is briefly specified in terms of applications targeted.

In any case, the web page advertising a given software also makes related publications available.

4.1.1. vorpatch and compatch: Modeling and Comparing Protein Binding Patches

Participants: Frédéric Cazals, Noël Malod-Dognin.

Context. Our work on the problem of modeling and comparing atomic resolution protein interfaces has been discussed in sections 5.4.1 and 5.1.1 The programs undertaking these two tasks are respectively named `vorpatch` and `compatch`.

Distribution. Binaries for `vorpatch` and `compatch` are available from http://cgal.inria.fr/abs/vorpatch-compatch/.

4.1.2. voratom: Modeling with Toleranced Models

Participants: Frédéric Cazals, Tom Dreyfus.

Context. Our TOleranced Model framework has been described in sections 5.2.1 and 5.2.2. The corresponding software package includes programs to (i) perform the segmentation of (probability) density maps, (ii) construct toleranced models, (iii) explore toleranced models (geometrically and topologically), (iv) compute Maximal Common Induced Sub-graphs (MCIS) and Maximal Common Edge Sub-graphs (MCES) to assess the pairwise contacts encoded in a TOM.

Distribution. Binaries for the aforementioned programs are made available from http://cgal.inria.fr/abs/voratom/.

4.1.3. wsheller: Selecting Water Layers in Solvated Protein Structures

Participants: Frédéric Cazals, Christine Roth.

Context. Given a snapshot of a molecular dynamics simulation, a classical problem consists of *quenching* that structure—minimizing the potential energy of the solute together with selected layers of solvent molecules. The program `wsheller` provides a solution to the water layer selection, and incorporates a topological control of the layers selected.

Distribution. Binaries for `wsheller` are available from http://cgal.inria.fr/abs/wsheller/.

4.1.4. intervor: Modeling Macro-molecular Interfaces

Participant: Frédéric Cazals.

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

Context. 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. Our 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.

Distribution. The following web site http://cgal.inria.fr/abs/Intervor serves two purposes: on the one hand, calculations can be run from the web site; on the other hand, binaries are distributed for Linux. To the best of our knowledge, this software is the only publicly available one for analyzing Voronoi interfaces in macro-molecular complexes.

4.1.5. vorlume: Computing Molecular Surfaces and Volumes with Certificates

Participant: Frédéric Cazals.

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

Context. 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, called `vorlume`. This program is based on 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.

Distribution. Binaries for `Vorlume` is available from http://cgal.inria.fr/abs/Vorlume.

4.1.6. ESBTL: theEasy Structural Biology Template Library

Participant: Frédéric Cazals.

In collaboration with S. Loriot (the Geometry Factory), and J. Bernauer, from the EPI AMIB.

Context. The ESBTL (Easy Structural Biology Template Library) is a lightweight C++ library that allows the handling of PDB data and provides a data structure suitable for geometric constructions and analyses.

Distribution. The source C++ code is available from http://esbtl.sourceforge.net/.

4.1.7. A_purva: Comparing Protein Structure 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.

Context. 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. The solver `A_purva` is 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.

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

5. New Results

5.1. Modeling Interfaces and Contacts

5.1.1. On the Morphology of Protein Binding Patches

Participants: Frédéric Cazals, Noël Malod-Dognin.

In collaboration with A. Bansal, former summer intern from IIT Bombay.

Understanding the specificity of protein interactions is a central question in structural biology, whence the importance of models for protein binding patches—a patch refers to the collection of atoms of a given partner accounting for the interaction. To improve our understanding of the relationship between the structure of binding patches and the biological function of protein complexes, we present a binding patch model decoupling the topological and geometric properties [21]. While the geometry is classically encoded by the 3D positions of the atoms, the topology is recorded in a graph encoding the relative position of concentric shells partitioning the interface atoms. The topological - geometric duality provides the basis of a generic dynamic programming based algorithm to compare patches, which is instantiated to respectively favor topological or geometric comparisons.

On the biological side, using a dataset of 92 co-crystallized structures organized in biological sub-families, we exploit our encoding and the two comparison algorithms in two directions. First, we show that Nature enjoyed the topological and geometric degrees of freedom independently while retaining a finite set of qualitatively distinct topological signatures, and show that topological similarity is a less stringent notion that the ubiquitously used geometric similarity. Second, we analyze the topological and geometric coherence of binding patches within sub-families and across the whole database, and show that complexes related to the same biological function can encompass geometrically distinct shapes. Previous work on binding patches focused on the investigation of correlations between structural parameters and biochemical properties on the one hand, and on structural comparison algorithms on the other hand. We believe that the abstraction coded by the topological - geometric duality paves the way to new classifications, in particular in the context of flexible docking.

The corresponding software is presented in section 4.1.1.

5.2. Modeling Macro-molecular Assemblies

5.2.1. Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models

Participants: Frédéric Cazals, Tom Dreyfus.

In collaboration with Valérie Doye, Institut Jacques Monod, Paris.

In [20], we introduce TOleranced Models (TOM), a generic and versatile framework meant to handle models of macro-molecular assemblies featuring uncertainties on the shapes and the positions of proteins. A TOM being a continuum of nested shapes, the inner (resp. outer) ones representing high (low) confidence regions, we present statistics to assess features of this continuum at multiple scales. While selected statistics target topological aspects (pairwise contacts, complexes involving proteins of prescribed types), others are of geometric nature (geometric accuracy of complexes). We validate the TOM framework on recent average models of the Nuclear Pore Complex (NPC) obtained from reconstruction by data integration, and confront our statistics against experimental findings related to sub-complexes of the NPC. In a broader perspective, the TOM framework should prove instrumental to handle uncertainties of various kind, in particular in electron-microscopy and crystallography.

5.2.2. Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Sub-complexes

Participants: Frédéric Cazals, Tom Dreyfus.

Reconstruction by data integration is an emerging trend to reconstruct large protein assemblies, but uncertainties on the input data yield average models whose quantitative interpretation is challenging. This paper presents methods to probe fuzzy models of large assemblies against atomic resolution models of sub-systems.

Consider a Toleranced Model (TOM) of a macro-molecular assembly, namely a continuum of nested shapes representing the assembly at multiple scales. Also consider a template namely an atomic resolution 3D model of a sub-system of this assembly—also called a complex. We present algorithms performing a multi-scale assessment of the complexes of the TOM, by comparing the pairwise contacts which appear in the TOM against those of the template. These operations reduce to the comparison of graphs, which we perform by computing Maximal Common Induced Sub-graphs (MCIS) and Maximal Common Edge Sub-graphs (MCES).

We apply this machinery to recent average models of the NPC. First, we show how our contact analysis allows assessing the quality of probability density maps. Regarding particular sub-systems of the NPC, we focus on the Y-complex and on the T-complex. In particular for the latter, our analysis suggests a new 3D template of pairwise contacts.

We believe that these tools should become standard to assess the reconstruction of fuzzy assemblies.

The software associated to these developments is presented in section 4.1.2.

5.3. Protein Shape Matching and Family Identification

5.3.1. Using Dominances for Solving the Protein Family Identification Problem

Participant: Noël Malod-Dognin.

In collaboration with R. Andonov (IRISA), M. Le Boudic-Jamin (IRISA) and P. Kamath (former summer intern within the SYMBIOSE *project at IRISA).*

The 3D structure of macro-molecules underpins all biological functions. Similarities between protein structures may come from evolutionary relationships, and similar protein structures relate to similar functions.

The exponential growth of the number of known protein structures in the Protein Data Bank over the past decade led to the problem of protein classification. We mean here how to automatically insert new protein structures into an already existing classified database $\mathcal{Q} = \{q_1, q_2, \cdots, q_m\}$ such as CATH or SCOP. The problem of determining in which classes new structures $\mathcal{P}$=$\{p_1, p_2, \cdots, p_n\}$ belong, according to a similarity function $S : \mathcal{Q} \times \mathcal{P} \rightarrow \mathcal{R}^+$, is referred here as the Protein Family Identification Problem (FIP).

There are computational pitfalls in the FIP . The number of similarity scores $S(q_i, p_j)$ that need to be computed is $|\mathcal{Q}| \times |\mathcal{P}|$, where $|\mathcal{P}|$ can be very large (there are currently 152920 classified protein structures in the expert classification CATH). Moreover, computing a single similarity score is often equivalent to solving a NP-hard problem (ex: DALI, DAST, CMO, VAST, etc...).

In [17] and [18], we propose a notion of dominance between the protein structure comparison instances that allows the computation of optimal FIP without optimally solving all the comparison instances, and thus reduces the effect of the NP-Hardness of the similarity score.

5.4. Algorithmic Foundations

5.4.1. Shape Matching by Localized Calculations of Quasi-isometric Subsets

Participants: Frédéric Cazals, Noël Malod-Dognin.

Consider a protein complex involving two partners, the receptor and the ligand. In [16], we address the problem of comparing their binding patches, i.e. the sets of atoms accounting for their interaction. This problem has been classically addressed by searching quasi-isometric subsets of atoms within the patches, a task equivalent to a maximum clique problem, a NP-hard problem, so that practical binding patches involving up to 300 atoms cannot be handled. We extend previous work in two directions. First, we present a generic encoding of shapes represented as cell complexes. We partition a shape into concentric shells, based on the shelling order of the cells of the complex. The shelling order yields a shelling tree encoding the geometry and the topology of the shape. Second, for the particular case of cell complexes representing protein binding patches, we present three novel shape comparison algorithms. These algorithms combine a Tree Edit Distance calculation (TED) on shelling trees, together with Edit operations respectively favoring a topological or a geometric comparison of the patches. We show in particular that the geometric TED calculation strikes a balance, in terms of accuracy and running time between a purely geometric and topological comparisons, and we briefly comment on the biological findings reported in a companion paper [21].

6. Dissemination

6.1. Animation of the scientific community

6.1.1. Conference Program Committees

– F. Cazals was member of the following PC:

- Symposium on Geometry Processing.
- SIAM Conference on Geometric and Physical Modeling.
- International conference on Pattern Recognition in Bioinformatics.

6.1.2. Ph.D. thesis and HDR Committees

– F.Cazals acted as *rapporteur* of the following habilitation defense:

- Dave Ritchie, University of Nancy, April 2011, *Rapporteur*. Habilitation memoir on *High performance algorithms for molecular shape recognition*.

6.1.3. Appointments

– F. Cazals is member of the scientific committee of *GDR Bio-informatique-Moléculaire*, in charge of activities related to computational structural biology.

– F. Cazals is member of the scientific committee of the exposition *Leonard de Vinci: la Nature et l'Invention*, Cité des Sciences.

– Until September 2011, F. Cazals was coordinating, together with Pierre Kornprobst, the Master of Science in Computational Biology and Medicine, http://cbb.unice.fr.

6.2. Teaching

(Master) Ecole Centrale Paris, France, 3rd year of the engineering curriculum in applied mathematics. Course on *Geometric and topological modeling with applications in biophysics*, taught by F. Cazals (24h).

(Master) Université de Nice Sophia Antipolis, France, Master of Science in Computational Biology (second year). Course on *Algorithmic Problems in Computational Structural Biology*, taught by F. Cazals (24h).

(PhD thesis, defended) T. Dreyfus, *Assessing the Reconstruction of Macro-molecular Assemblies: the Example of the Nuclear Pore Complex*, Université de Nice Sophia Antipolis, defended on December the 20th. Advisor: F. Cazals.

(PhD thesis, ongoing) C. Roth, *Modeling the flexibility of macro-molecules: theory and applications*, Université de Nice Sophia Antipolis. Advisor: F. Cazals.

(PhD thesis, ongoing) A. Lheritier, *Scoring and discriminating in high-dimensional spaces: a geometric based approach of statistical tests*, Université de Nice Sophia Antipolis. Advisor: F. Cazals.

(PhD thesis, ongoing) D. Agarwal, *Towards nano-molecular design: advanced algorithms for modeling large protein assemblies*, Univ. of Nice - Sophia-Antipolis. Advisor: F. Cazals.

7. 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

Doctoral Dissertations and Habilitation Theses

[11] T. DREYFUS. *Assessing the Reconstruction of Macro-molecular Assemblies: the Example of the Nuclear Pore Complex*, Informatique, Université de Nice Sophia Antipolis, 2011.

Articles in International Peer-Reviewed Journal

[12] R. ANDONOV, N. MALOD-DOGNIN, N. YANEV. *Maximum Contact Map Overlap Revisited*, in "Journal of Computational Biology", January 2011, vol. 18, n^o 1, p. 1-15 [*DOI :* 10.1089/CMB.2009.0196], http://hal.inria.fr/inria-00536624/en.

[13] F. CAZALS, D. COHEN-STEINER. *Reconstructing 3D compact sets*, in "Computational Geometry Theory and Applications", 2011, vol. 45, n^o 1-2, p. 1–13.

[14] F. CAZALS, H. KANHERE, S. LORIOT. *Computing the Volume of Union of Balls: a Certified Algorithm*, in "ACM Transactions on Mathematical Software", 2011, vol. 38, n^o 1, p. 1–20.

[15] S. LORIOT, S. SACHDEVA, K. BASTARD, C. PREVOST, F. CAZALS. *On the Characterization and Selection of Diverse Conformational Ensembles*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2011, vol. 8, n^o 2, p. 487–498.

International Conferences with Proceedings

[16] F. CAZALS, N. MALOD-DOGNIN. *Shape Matching by Localized Calculations of Quasi-isometric Subsets, with Applications to the Comparison of Protein Binding Patches*, in "The 6th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB)", Delft, Netherlands, LOOG, MARCO, WESSELS, LODEWYK, REINDERS, MARCEL, D. RIDDER, DICK (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, June 2011, vol. 7036, p. 272-283, http://hal.inria.fr/inria-00603375/en.

[17] N. MALOD-DOGNIN, M. L. BOUDIC-JAMIN, P. KAMATH, R. ANDONOV. *Using Dominances for Solving the Protein Family Identification Problem*, in "11th Workshop on Algorithms in Bioinformatics (WABI 2011)", Saarbrücken, Germany, PRZYTYCKA, TERESA, SAGOT, MARIE-FRANCE (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, July 2011, vol. 6833, p. 201-212, Published in Workshop on Algorithms for Bioinformatics (WABI 2011) http://predictioncenter.org [*DOI :* 10.1007/978-3-642-23038-7_18], http://hal.inria.fr/inria-00609432/en.

National Conferences with Proceeding

[18] M. L. Boudic-Jamin, N. Malod-Dognin, A. Cornu, J. Nicolas, R. Andonov. *Identification rapide de familles protéiques par dominance*, in "12th Annual Congress of the French National Society of Operations Research and Decision Science (ROADEF)", Saint-Étienne, France, École Nationale Supérieure des Mines de Saint-Étienne, March 2011, vol. 2, p. 791-792, Publié dans le douzième congrès de la Société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF 2011)., http://hal.inria.fr/inria-00611457/en.

Research Reports

[19] F. Cazals, T. Dreyfus. *Assessing the Reconstruction of Macro-molecular Assemblies: the Example of the Nuclear Pore Complex*, INRIA, January 2011, n^o RR-7513, http://hal.inria.fr/inria-00559117/en.

[20] T. Dreyfus, V. Doye, F. Cazals. *Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models*, INRIA, 2011, n^o RR-7768, http://hal.inria.fr/inria-00635590/en.

[21] N. Malod-Dognin, A. Bansal, F. Cazals. *Characterizing the Morphology of Protein Binding Patches*, INRIA, September 2011, n^o RR-7743, http://hal.inria.fr/inria-00626548/en.

References in notes

[22] 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.

[23] 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.

[24] F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali. *Integrating Diverse Data for Structure Determination of Macromolecular Assemblies*, in "Ann. Rev. Biochem.", 2008, vol. 77, p. 11.1–11.35.

[25] O. Becker, A. D. Mackerell, B. Roux, M. Watanabe. *Computational Biochemistry and Biophysics*, M. Dekker, 2001.

[26] A.-C. Camproux, R. Gautier, P. Tuffery. *A Hidden Markov Model derived structural alphabet for proteins*, in "J. Mol. Biol.", 2004, p. 591-605.

[27] 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.

[28] 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.

[29] D. Cohen-Steiner, H. Edelsbrunner, J. Harer. *Stability of Persistence Diagrams*, in "ACM SoCG", 2005.

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

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

[32] 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.

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

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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