
Activity Report 2012

# 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

*Creation of the Project-Team:* July 01, 2008 .

# 1. Members

**Research Scientist**

Frédéric Cazals [Team leader; DR2 Inria, HdR]

**PhD Students**

Deepesh Agarwal [ INRIA ]
Alix Lhéritier [INRIA CORDI-S fellow]
Christine Roth [INRIA CORDI-S fellow]
Tom Dreyfus [Engineer, from the 09/01/2012]

**Post-Doctoral Fellow**

Noël Malod-Dognin [INRIA, until 06/30/2012]

**Administrative Assistant**

Nelly Bessega [Assistant of ABS and GEOMETRICA]

# 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 [46]. 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 circa 75,000 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 [44], [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) [35], 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.

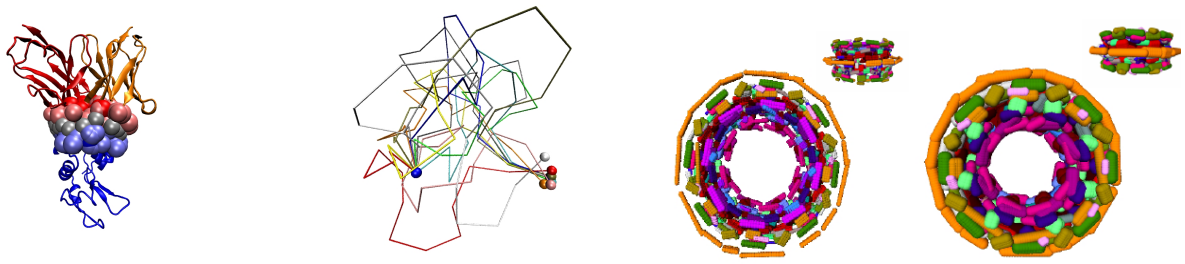
## 2.2. Highlights of the Year

Three key achievements were obtained in 2012.

The first one deals with the problem of modeling high resolution protein complexes, a topic for which we came up with an original binding patch model [14]. Our model not only provides more accurate descriptors of key quantities (the binding affinity in particular), but also sheds new light on the flexibility of proteins upon docking. These developments will in particular be used to investigate complexes from the immune system in the future.

The second one deals with the problem of modeling large protein assemblies, involving up to hundreds of polypeptide chains. We finalized the application of our Toleranced Models framework to the nuclear pore complex [13], [19], and started to produce novel algorithms for mass-spectrometry data [18], an emerging technique to infer structural information on large molecular machines.

Finally, we have also made a steady progress on algorithmic foundations, in particular on the problem of developing a Morse theory for point cloud data, in the perspective of analyzing molecular dynamics data. Tests are currently on the way, so that this work will be advertised in 2013.



(a) (b) (c)

*Figure 1.* ***Geometric constructions in computational structural biology.*** *(a) An antibody-antigen complex, with interface atoms identified by our Voronoi based interface model [10], [2]. This model is instrumental in mining correlations between structural and biological as well as biophysical properties of protein complexes. (b) A diverse set of conformations of a backbone loop, selected thanks to a geometric optimization algorithm [11]. Such conformations are used by mean field theory based docking algorithms. (c) A toleranced model (TOM) of the nuclear pore complex, visualized at two different scales [13]. The parameterized family of shapes coded by a TOM is instrumental to identify stable properties of the underlying macro-molecular system.*

# 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

Docking, interfaces, protein complexes, structural alphabets, scoring functions, Voronoi diagrams, arrangements of balls

**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 [46]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [49]. 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 [43]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [38].

[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.

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 [47], [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 [39]. 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 [48], 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 $C_\alpha$ 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 [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 [42]. 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

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

### *3.3.1. Reconstruction by data integration*

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [22]. In a nutshell, the strategy is reminiscent

[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.

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 [21], 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 [20], 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 [20]. 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

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

**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.

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

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) [45]. 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 [41], to Morse theory [36] and to analysis of meta-stable states of time series [37] 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 [40].

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 [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].

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 2012 is briefly described as the context is presented in the sections dedicated to new results. On the other hand, the software made available before 2012 is briefly specified in terms of applications targeted.

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

### *4.1.1. addict: Stoichiometry Determination for Mass Spectrometry Data*

**Participants:** Deepesh Agarwal, Frédéric Cazals, Noël Malod-Dognin.

**Context.** Our work on the stoichiometry determination (SD) problem for noisy data in structural proteomics is described in section 5.2.1. The `addict` software suite not only implements our algorithms `DP++` and `DIOPHANTINE`, but also important algorithms to determine the so-called Frobenius number of a vector of protein masses, and also to estimate the number of solutions of a SD problem, from an unbounded knapsack problem.

**Distribution.** Binaries for the `addict` software suite are made available from http://team.inria.fr/abs/software/voratom/.

### *4.1.2. vorpatch and compatch: Modeling and Comparing Protein Binding Patches*

**Participants:** Frédéric Cazals, Noël Malod-Dognin.

**Context.** Modeling protein binding patches is a central problem to foster our understanding of the stability and of the specificity of macro-molecular interactions. We developed a binding patch model which encodes morphological properties, allows an atomic-level comparison of binding patches at the geometric and topological levels, and allows estimating binding affinities—with state-of-the-art results on the protein complexes of the binding affinity benchmark. Given a protein complex, `vorpatch` compute the binding patches, while the program `compatch` allows comparing two patches.

**Distribution.** Binaries for `VORPATCH` and `COMPATCH` are available from http://team.inria.fr/abs/software/vorpatch-compatch.

### *4.1.3. voratom: Modeling Protein Assemblies with Toleranced Models*

**Participants:** Frédéric Cazals, Tom Dreyfus.

**Context.** 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. Modeling such assemblies is especially challenging due to their plasticity (the proteins involved may change along the cell cycle), their size, and also the flexibility of the sub-units. To cope with these difficulties, a reconstruction strategy known as Reconstruction by Data Integration (RDI), aims at integrating diverse experimental data. But the uncertainties on the input data yield equally uncertain reconstructed models, calling for quantitative assessment strategies.

To leverage thee reconstruction results, we introduced TOleranced Model (TOM) framework, which inherently accommodates uncertainties on the shape and position of proteins. The corresponding software package, `VORATOM`, 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 software package `VORATOM` are made available from http://team.inria.fr/abs/software/voratom/.

### *4.1.4. 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://team.inria.fr/abs/software/wsheller.

### *4.1.5. intervor: Modeling Macro-molecular Interfaces*

**Participant:** Frédéric Cazals.

*In collaboration with S. Loriot (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 website http://team.inria.fr/abs/software/intervor serves two purposes: on the one hand, calculations can be run from the website; on the other hand, binaries are made available. To the best of our knowledge, this software is the only publicly available one for analyzing Voronoi interfaces in macromolecular complexes.

#### 4.1.6. *vorlume: Computing Molecular Surfaces and Volumes with Certificates*

**Participant:** Frédéric Cazals.

*In collaboration with S. Loriot (The* GEOMETRY FACTORY*, France)*

**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://team.inria.fr/abs/software/vorlume.

#### 4.1.7. *ESBTL: theEasy Structural Biology Template Library*

**Participant:** Frédéric Cazals.

*In collaboration with S. Loriot (The* GEOMETRY FACTORY*, France) and J. Bernauer (Inria AMIB, France).*

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

#### 4.1.8. *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.orghttp://apurva.genouest.org.

# 5. New Results

## 5.1. Modeling Interfaces and Contacts

Docking, scoring, interfaces, protein complexes, scoring functions, Voronoi diagrams, arrangements of balls.

#### 5.1.1. *Modeling Macro-molecular Complexes : a Journey Across Scales*

**Participants:** Frédéric Cazals, Tom Dreyfus.

*In collaboration with C. Robert (IBPC / CNRS, Paris, France).*

While proteins and nucleic acids are the fundamental components of an organism, Biology itself is based on the interactions they make with each other. Analyzing macromolecular interactions typically requires handling systems involving from two to hundreds of polypeptide chains. After a brief overview of the modeling challenges faced in computational structural biology, the text [16] reviews concepts and tools aiming at improving our understanding of the link between the static structures of macromolecular complexes and their biophysical/biological properties. We discuss geometrical approaches suited to atomic-resolution complexes and to large protein assemblies; for each, we also present examples of their successful application in quantifying and interpreting biological data. This methodology includes state-of-the-art geometric analyses of surface area, volume, curvature, and topological properties (isolated components, cavities, voids, cycles) related to Voronoi constructions in the context of structure analysis. On the applied side, we present novel insights into real biological problems gained thanks to these modeling tools.

### *5.1.2. CSA: Comprehensive Comparison of Pairwise Protein Structure Alignments*

**Participant:** Noël Malod-Dognin.

*In collaboration with I. Wohlers (CWI / VU University Amsterdam, Netherlands), R. Andonov (Irisa / Rennes University, France), G.W. Klau (CWI / VU University Amsterdam, Netherlands).*

Protein structural alignment is a key method for answering many biological questions involving the transfer of information from well-studied proteins to less well-known proteins. Since structures are more conserved during evolution than sequences, structural alignment allows for the most precise mapping of equivalent residues. Many structure-based scoring schemes have been proposed and there is no consensus on which scoring is the best. Comparative studies also show that alignments produced by different methods can differ considerably. Based on the alignment engine derived from A_purva, we designed CSA (Comparative Structural Alignment), the first web server for computation, evaluation and comprehensive comparison of pairwise protein structure alignments at single residue level [15]. It offers the exact computation of alignments using the scoring schemes of DALI, Contact Map Overlap (CMO), MATRAS and PAUL. In CSA, computed or uploaded alignments can be explored in terms of many inter-residue distances, RMSD, and sequence-based scores. Intuitive visualizations also help in grasping the agreements and differences between alignments. The user can thus make educated decisions about the structural similarity of two proteins and, if necessary, post-process alignments by hand. CSA is available at http://csa.project.cwi.nl.

Upon publication [15], CSA was selected by *Nucleic Acids Research* as featured article of July 2012 (top 5% of papers in terms of originality, significance and scientific excellence.

## 5.2. Modeling Macro-molecular Assemblies

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

### *5.2.1. Stoichiometry Determination for Mass-spectrometry Data: the Interval Case*

**Participants:** Deepesh Agarwal, Frédéric Cazals, Noël Malod-Dognin.

In structural proteomics, given the individual masses of a set of protein types and the exact mass of a protein complex, the *exact stoichiometry determination problem (SD)*, also known as the money-change problem, consists of enumerating all the stoichiometries of these types which allow to recover the target mass. If the target mass suffers from experimental uncertainties, the *interval SD problem* consists of finding all the stoichiometry vectors compatible with a target mass within an interval.

We make contributions in two directions [18]. From a theoretical standpoint, we present a constant-memory space algorithm (`DIOPHANTINE`) and an output sensitive dynamic programming based algorithm (`DP++`), both inherently addressing the interval SD problem. From an applied perspective, we raise three points. First, we show that `DIOPHANTINE` and `DP++` yield an improvement from 3 to 4 orders of magnitude over state-of-the-art exact SD algorithms, for typical protein complexes facing uncertainties on the target mass in the range 0.1-1%. Second, we show that `DIOPHANTINE` behaves like an output-sensitive algorithm—especially when the interval

width increases, albeit such a property cannot be expected in general. Third, from a biological perspective, using a panel of biological complexes (eukaryotic translation factor, yeast exosome, 19S proteasome sub-unit, nuclear pore complex), we stress the importance of enumeration, even at a null noise level.

The programs accompanying this paper are available from http://team.inria.fr/abs/addict/.

## 5.3. Algorithmic Foundations

Voronoi diagrams, $\alpha$-shapes,

The work undertaken in this vein in 2012 will be finalized in 2013.

## 5.4. Immunology

Immune response, infection, antibodies, complementarity determining region (CDR)

### *5.4.1. Teleost Fish Mount Complex Clonal IgM and IgT Responses in Spleen Upon Systemic Viral Infection*

**Participant:** Frédéric Cazals.

*In collaboration with*

- *R. Castro, L. Journeau, A. Benmansour and P. Boudinot (INRA Jouy-en-Josas, France)*
- *H.P. Pham and A. Six (Univ. of Paris VI, France)*
- *O. Bouchez (INRA Castanet Tolosan, France)*
- *V. Giudicelli and M-P. Lefranc (IMGT / CNRS, Montpellier, France)*
- *E. Quillet (INRA Jouy-en-Josas, France)*
- *S. Fillatreau (Leibniz Institute, Berlin, Germany)*
- *O. Sunyer (Univ. of Pennsylvania, USA)*

Upon infection, B-lymphocytes expressing antibodies specific for the intruding pathogen develop clonal responses triggered by pathogen recognition via the B-cell receptor. The constant region of antibodies produced by such developing clones dictates their functional properties. In teleost fish, the clonal structure of B-cell responses and the respective contribution of the three isotypes IgM, IgD, and IdT remains unknown. The expression of IgM and IgT are mutually exclusive, leading to the existence of two B-cell subsets expressing either both IgM and IgD or only IgT. In [12], we undertook a comprehensive analysis of the variable heavy chain (VH) domain repertoires of the IgM, IgD and IgT in spleen of homozygous isogenic rainbow trout (Onchorhynus mykiss), before and after challenge with a rhabdovirus, the Viral Hemorrhagic Septicemia Virus (VHSV), using CDR3-length spectratyping and pyrosequencing of immunoglobulin (Ig) transcripts. In healthy fish, we observed distinct repertories for IgM, IgD and IgT respectively, with a few amplified $\mu$ and $\tau$ junctions, suggesting the presence of IgM and IgT secreting cells in the spleen. In infected animals, we detected complex and highly diverse IgM responses involving all VH subgroups, and dominated by a few large public and private clones. A lower number of robust clonal responses involving only a few VH were detected for the mucosal IgT, indicating that both $IgM^+$ and $IgT^+$ spleen B cells responded to systemic infection but at different degrees. In contrast, the IgD response to the infection was faint. Although the IgD and IgT present different structural features and evolutionary origin compared to mammalian IgD and IgA respectively, their implication in the B-cell response evokes these mouse and human counterparts. Thus, it appears that the general properties of antibody responses were already in place in common ancestors of fish and mammals, and were globally conserved during evolution with possible functional convergences.

# 6. Partnerships and Cooperations

## 6.1. National Initiatives

### *6.1.1. Projets Exploratoires Pluridisciplinaires from CNRS/Inria/INSERM*

Reconstruction by Data Integration (RDI) is an emerging paradigm to reconstruct large protein assemblies, as discussed in section 4.1.3.

Elaborating on our Toleranced Models framework, a geometric framework aiming at inherently accommodating uncertainties on the shapes and positions of proteins within large assemblies, we ambition within the scope of the two year long PEPS project entitled *Modeling Large Protein Assemblies with Toleranced Models* to (i) design TOM compatible with the flexibility of proteins, (ii) develop graph-based analysis of TOM, and (iii) perform experimental validations on the NPC.

## 6.2. European Initiatives

### *6.2.1. FP7 Projet*

#### *6.2.1.1. CG-Learning*

Title: Computational Geometric Learning (CGL)

Type: COOPERATION (ICT)

Defi: FET Open

Instrument: Specific Targeted Research Project (STREP)

Duration: November 2010 - October 2013

Coordinator: Friedrich-Schiller-Universität Jena (Germany)

Others partners: Jena Univ. (coord.), Inria (Geometrica Sophia, Geometrica Saclay, ABS), Tech. Univ. of Dortmund, Tel Aviv Univ., Nat. Univ. of Athens, Univ. of Groningen, ETH Zürich, Freie Univ. Berlin.

See also: http://cglearning.eu/http://cglearning.eu/

Abstract: *The Computational Geometric Learning project aims at extending the success story of geometric algorithms with guarantees to high-dimensions. This is not a straightforward task. For many problems, no efficient algorithms exist that compute the exact solution in high dimensions. This behavior is commonly called the curse of dimensionality. We try to address the curse of dimensionality by focusing on inherent structure in the data like sparsity or low intrinsic dimension, and by resorting to fast approximation algorithms.*

## 6.3. International Research Visitors

### *6.3.1. Internships*

- From May to July 2012, summer internship from Pratik Kumar (Indian Institute of Technology of Bombay). Topic: Modeling density maps in cryo electron microscopy.

# 7. Dissemination

## 7.1. Scientific Animation

### *7.1.1. Conference Program Committees*

– F. Cazals was member of the following PC:

- Symposium on Geometry Processing.
- Geometric Modeling and processing.
- ACM Symposium on Solid and Physical Modeling.
- IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.
- International conference on Pattern Recognition in Bioinformatics.

### 7.1.2. Appointments

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

### 7.1.3. Book

Having initiated and coordinated the Master of Science in Computational Biology, see http://cbb.unice.fr, F. Cazals and P. Kornprobst edited a book entitled *Modeling in Computational Biology and Medicine: A Multidisciplinary Endeavor* [17], with one chapter per class taught in this program.

## 7.2. Teaching - Supervision - Juries

### 7.2.1. 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)** University of Nice Sophia Antipolis, France, Master of Science in Computational Biology (http://cbb.unice.fr). Course on *Algorithmic problems in computational structural biology*, taught by F. Cazals (24h).

**(Winter school Algorithms in Structural Bio-informatics)** Together with J. Cortès from LAAS / CNRS (Toulouse), F. Cazals is organizing the winter school *Algorithms in Structural Bio-informatics* [4]. The goal of this winter school is to present state-of-the-art concepts, algorithms and software tools meant to analyze and predict macro-molecular assemblies, with a focus on methodological developments. We have accepted 25 students from all over the world.

### 7.2.2. Supervision

PhD & HdR:

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

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

# 8. Bibliography

## Major publications by the team in recent years

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

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

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

[4] http://www-sop.inria.fr/manifestations/algoSB/

[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 $\alpha$-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.

[11] 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.

# Publications of the year

## Articles in International Peer-Reviewed Journals

[12] R. CASTRO, L. JOURNEAU, H. PHAM, O. BOUCHEZ, V. GIUDICELLI, M.-P. LEFRANC, E. QUILLET, A. BENMANSOUR, F. CAZALS, A. SIX, S. FILLATREAU, O. SUNYER, P. BOUDINOT. *Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection*, in "PLOS Pathogens", 2012, In press.

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

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

[15] I. WOHLERS, N. MALOD-DOGNIN, R. ANDONOV, G. KLAU. *CSA: Comprehensive comparison of pairwise protein structure alignments*, in "Nucleic Acids Research", 2012, vol. 40, n$^{o}$ W1, p. W303–W309.

## Scientific Books (or Scientific Book chapters)

[16] F. CAZALS, T. DREYFUS, C. ROBERT. *Modeling Macro-molecular Complexes : a Journey Across Scales*, in "Modeling in Computational Biology and Medicine: A Multidisciplinary Endeavor", F. CAZALS, P. KORNPROBST (editors), Springer, 2013.

### Books or Proceedings Editing

[17] F. CAZALS, P. KORNPROBST (editors). *Modeling in Computational Biology and Medicine: A Multidisciplinary Endeavor*, Springer, 2013, 315.

### Research Reports

[18] D. AGARWAL, F. CAZALS, N. MALOD-DOGNIN. *Stoichiometry Determination for Mass-spectrometry Data: the Interval Case*, Inria, October 2012, n$^{o}$ RR-8101, 31, http://hal.inria.fr/hal-00741491.

[19] T. DREYFUS, V. DOYE, F. CAZALS. *Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Complexes*, Inria, October 2012, n$^{o}$ RR-8118, 20, http://hal.inria.fr/hal-00745558.

## References in notes

[20] F. ALBER, S. DOKUDOVSKAYA, L. VEENHOFF, W. ZHANG, J. KIPPER, D. DEVOS, A. SUPRAPTO, O. KARNI-SCHMIDT, R. WILLIAMS, B. CHAIT, M. ROUT, A. SALI. *Determining the architectures of macromolecular assemblies*, in "Nature", Nov 2007, vol. 450, p. 683-694.

[21] F. ALBER, S. DOKUDOVSKAYA, L. VEENHOFF, W. ZHANG, J. KIPPER, D. DEVOS, A. SUPRAPTO, O. KARNI-SCHMIDT, R. WILLIAMS, B. CHAIT, A. SALI, M. ROUT. *The molecular architecture of the nuclear pore complex*, in "Nature", 2007, vol. 450, n$^{o}$ 7170, p. 695–701.

[22] 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.

[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 compact sets in Euclidean space*, in "Discrete and Computational Geometry", 2009, vol. 41, n$^{o}$ 3, p. 461–479.

[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 SoCG", 2005.

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

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

[30] 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.

[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)", M. G. ROSSMANN, E. ARNOLD (editors), Springer, 2001, p. 531–539.

[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.", 2001, vol. 11, p. 231-235.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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