Activity Report 2013

Project-Team ABS

Algorithms, Biology, Structure

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Computational Biology

Table of contents

Project-Team ABS

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

Creation of the Project-Team: 2008 July 01.

1. Members

Research Scientist
Frédéric Cazals [Team leader, Inria, Senior Researcher, HdR]

External Collaborator
Charles Robert [CNRS, HdR]

Engineer
Tom Dreyfus [Inria]

PhD Students
Deepesh Agarwal [Inria]
Alix Lhéritier [Inria]
Simon Marillet [INRA, from Oct 2013]
Christine Roth [Inria, granted by FP7 CG Learning project]

Administrative Assistant
Florence Barbara [Inria, from Jul 2013]

Other
Angeliki Kalamara [Inria, Summer Intern, from March 2013 until July 2013]

2. Overall Objectives

2.1. Overall Objectives

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 [43]. 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, the order of magnitude of the number of genomes sequenced is of the order of one thousand, which results in circa one million of genes recorded in the manually curated Swiss-Prot database. On the other hand, the Protein Data Bank contains circa 90,000 structures. Thus, the paucity of structures with respect to the known number of genes calls for modeling in structural biology, so as to foster our understanding of the structure-to-function relationship.

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 [41], [30] and later Connolly [26], while defining molecular surfaces and developing algorithms to compute them, established the connexions between molecular modeling and geometric constructions. Remarkably, a number of difficult problems (e.g. additively weighted Voronoi diagrams) were touched upon in these early days.

The models developed in this vein are instrumental in investigating the interactions of molecules for which no structural data is available. But such models often fall short from providing complete answers, which we illustrate with the folding problem. On one hand, as the conformations of side-chains belong to discrete sets (the so-called rotamers or rotational isomers) [32], 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, while time-steps used in molecular dynamics simulations are of the order of the femto-second, so that biologically relevant time-scales are out 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 .

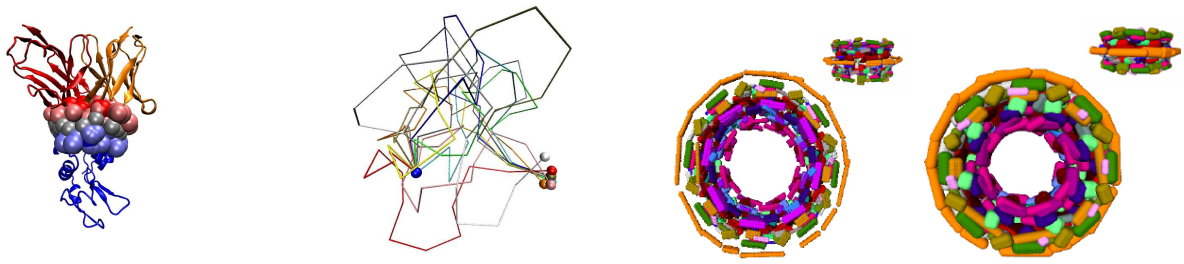
3. Research Program

3.1. Introduction

The research conducted by ABS focuses on three main directions in Computational Structural Biology (CSB), together with the associated methodological developments:
– Modeling interfaces and contacts,
– Modeling macro-molecular assemblies,
– Modeling the flexibility of macro-molecules,
– Algorithmic foundations.

3.2. Modeling Interfaces and Contacts

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



(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 [8], [1]. 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 [10]. 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 [9]. The parameterized family of shapes coded by a TOM is instrumental to identify stable properties of the underlying macro-molecular system.*

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 [43]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [46]. Current investigations follow two routes. From the experimental perspective [29], 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 [40]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [35].

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 [24], 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 [44], [31]. 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 [36]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

[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.

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 [8]. 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 [25]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [45], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond —a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [28].

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 [39]. 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

Keywords: 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 [23]. 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 [22], 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 [21], 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 [21]. 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.

3.4. Modeling the Flexibility of Macro-molecules

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

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 generate ensembles of conformations, called `conformers`, and so do molecular dynamics (MD) simulations. 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.

At the side-chain level, the question of improving rotamer libraries is still of interest [27]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [42]. 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 [38], to Morse theory [33] and to analysis of meta-stable states of time series [34] have been proposed.

3.5. Algorithmic Foundations

Keywords: Computational geometry, computational topology, optimization, data analysis.

Making a stride towards a better understanding of the biophysical questions discussed in the previous sections requires various methodological developments, which we briefly discuss now.

3.5.1. Modeling interfaces and contacts

In modeling interfaces and contacts, one may favor geometric or topological information.

On the geometric side, the problem of modeling contacts at the atomic level is tantamount to encoding multi-body relations between an atom and its neighbors. On the 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. The information gathered while modeling contacts can further be integrated into interface models.

On the topological side, one may favor constructions which remain stable if each atom in a structure *retains* the same neighbors, even though the 3D positions of these neighbors change to some extent. This process is observed in flexible docking cases, and call for the development of methods to encode and compare shapes undergoing tame geometric deformations.

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

3.5.2. Modeling macro-molecular assemblies

In dealing with large assemblies, a number of methodological developments are called for.

On the experimental side, of particular interest is the disambiguation of proteomics signals. For example, TAP and mass spectrometry data call for the development of combinatorial algorithms aiming at unraveling pairwise contacts between proteins within an assembly. Likewise, density maps coming from electron microscopy, which are often of intermediate resolution (5-10Å) call the development of noise resilient segmentation and interpretation algorithms. The results produced by such algorithms can further be used to guide the docking of high resolutions crystal structures into maps.

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, a process reminiscent from non convex optimization. The second one encompasses assessment methods, in order to single out the reconstructions which best comply with the experimental data. For that endeavor, the development of geometric and topological models accommodating uncertainties is particularly important.

3.5.3. Modeling the flexibility of macro-molecules

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 [37].

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

4. Application Domains

4.1. Structural Biology and Biophysics

As the name of the project-team suggest, *Algorithms-Biology-Structure* is primarily concerned with the investigation of the structure-to-function relationship in structural biology and biophysics.

5. Software and Platforms

5.1. Software

This section briefly comments on all the software distributed by ABS . On the one hand, the software released in 2013 is briefly described as the context is presented in the sections dedicated to new results. On the other hand, the software made available before 2013 is briefly specified in terms of applications targeted.

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

5.1.1. addict: Stoichiometry Determination from Mass Spectrometry Data

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

Context. Given the individual masses of the proteins present in a complex, together with the mass of that complex, *stoichiometry determination* (SD) consists of computing how many copies of each protein are needed to account for the overall mass of the complex. Our work on the stoichiometry determination (SD) problem for noisy data in structural proteomics is described in [17]. 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/addict/.

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

Participants: Frédéric Cazals, Noël Malod-Dognin.

Context. Modeling protein binding patches, i.e. the sets of atoms responsible of an interaction, 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 binary protein complex, `vorpatch` identifies the binding patches, and computes a topological encoding of each patch, defined as an *atom shelling tree* generalizing the core-rim model. The program `compatch` allows comparing two patches via the comparison of their atom shelling trees, by favoring either a geometric or a topological comparison.

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

5.1.3. 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 these reconstruction results, we introduced TOleranced Model (TOM) framework, which inherently accommodates uncertainties on the shape and position of proteins represented as density maps — maps from cryo electron-microscopy or maps stemming from reconstruction by data integration. In a TOM, a fuzzy molecule is sandwiched between two union of concentric balls, the size of the region between these two unions conveying information on the uncertainties.

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/.

5.1.4. 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 macro-molecular complexes.

5.1.5. 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.

5.1.6. ESBTL: the Easy 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, such as those proposed by `INTERVOR`, `VORPATCH` and `COMPATCH`.

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

6. New Results

6.1. Modeling Interfaces and Contacts

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

The work undertaken in this vein in 2013 will be finalized in 2014.

6.2. Modeling Macro-molecular Assemblies

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

6.2.1. Connectivity Inference in Mass Spectrometry based Structure Determination

Participants: Frédéric Cazals, Deepesh Agarwal.

In collaboration with J. Araujo, and C. Caillouet, and D. Coudert, and S. Pérennes, from the COATI project-team (Inria-CNRS).

In [14], we consider the following MINIMUM CONNECTIVITY INFERENCE problem (MCI), which arises in structural biology: given vertex sets $V_i \subseteq V, i \in I$, find the graph $G = (V, E)$ minimizing the size of the edge set E, such that the sub-graph of G induced by each V_i is connected. This problem arises in structural biology, when one aims at finding the pairwise contacts between the proteins of a protein assembly, given the lists of proteins involved in sub-complexes. We present four contributions.

First, using a reduction of the set cover problem, we establish that MCI is APX-hard. Second, we show how to solve the problem to optimality using a mixed integer linear programming formulation (`MILP`). Third, we develop a greedy algorithm based on union-find data structures (`Greedy`), yielding a $2(\log_2 |V| + \log_2 \kappa)$-approximation, with κ the maximum number of subsets V_i a vertex belongs to. Fourth, application-wise, we use the MILP and the greedy heuristic to solve the aforementioned connectivity inference problem in structural biology. We show that the solutions of `MILP` and `Greedy` are more parsimonious than those reported by the algorithm initially developed in biophysics, which are not qualified in terms of optimality. Since MILP outputs a set of optimal solutions, we introduce the notion of *consensus solution*. Using assemblies whose pairwise contacts are known exhaustively, we show an almost perfect agreement between the contacts predicted by our algorithms and the experimentally determined ones, especially for consensus solutions.

6.3. Algorithmic Foundations

Computational geometry, Computational topology, Voronoi diagrams, α-shapes, Morse theory.

6.3.1. Greedy Geometric Algorithms for Collection of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining

Participants: Frédéric Cazals, Tom Dreyfus.

In collaboration with S. Sachdeva (Princeton University, USA), and N. Shah (Carnegie Mellon University, USA).

Choosing balls to best approximate a 3D object is a non trivial problem. To answer it, in [18], we first address the *inner approximation* problem, which consists of approximating an object $\mathcal{F}_{\mathcal{O}}$ defined by a union of n balls with $k < n$ balls defining a region $\mathcal{F}_{\mathcal{S}} \subset \mathcal{F}_{\mathcal{O}}$. This solution is further used to construct an *outer approximation* enclosing the initial shape, and an *interpolated approximation* sandwiched between the inner and outer approximations.

The inner approximation problem is reduced to a geometric generalization of weighted max k-cover, solved with the greedy strategy which achieves the classical $1 - 1/e$ lower bound. The outer approximation is reduced to exploiting the partition of the boundary of $\mathcal{F}_{\mathcal{O}}$ by the Apollonius Voronoi diagram of the balls defining the inner approximation.

Implementation-wise, we present robust software incorporating the calculation of the exact Delaunay triangulation of points with degree two algebraic coordinates, of the exact medial axis of a union of balls, and of a certified estimate of the volume of a union of balls. Application-wise, we exhibit accurate coarse-grain molecular models using a number of balls 20 times smaller than the number of atoms, a key requirement to simulate crowded cellular environments.

6.3.2. Towards Morse Theory for Point Cloud Data

Participants: Frédéric Cazals, Christine Roth.

In collaboration with C. Robert (IBPC / CNRS, Paris, France), and C. Mueller (ETH, Zurich).

Morse theory provides a powerful framework to study the topology of a manifold from a function defined on it, but discrete constructions have remained elusive due to the difficulty of translating smooth concepts to the discrete setting.

Consider the problem of approximating the Morse-Smale (MS) complex of a Morse function from a point cloud and an associated nearest neighbor graph (NNG). While following the constructive proof of the Morse homology theorem, we present novel concepts for critical points of any index, and the associated Morse-Smale diagram [19].

Our framework has three key advantages. First, it requires elementary data structures and operations, and is thus suitable for high-dimensional data processing. Second, it is gradient free, which makes it suitable to investigate functions whose gradient is unknown or expensive to compute. Third, in case of under-sampling and even if the exact (unknown) MS diagram is not found, the output conveys information in terms of ambiguous flow, and the Morse theoretical version of topological persistence, which consists in canceling critical points by flow reversal, applies.
On the experimental side, we present a comprehensive analysis of a large panel of bi-variate and tri-variate Morse functions whose Morse-Smale diagrams are known perfectly, and show that these diagrams are recovered perfectly.

In a broader perspective, we see our framework as a first step to study complex dynamical systems from mere samplings consisting of point clouds.

6.4. Misc

Computational Biology, Biomedicine.

6.4.1. Book

Participant: Frédéric Cazals.

Edited in collaboration with P. Kornprobst, from the Neuromathcomp project-team.

Biology and biomedicine currently undergo spectacular progresses due to a synergy between technological advances and inputs from physics, chemistry, mathematics, statistics and computer science. The goal of the book [15] is to evidence this synergy, by describing selected developments in the following fields: bioinformatics, biomedicine, neuroscience.

This book is unique in two respects. First, by the variety and scales of systems studied. Second, by its presentation, as each chapter presents the biological or medical context, follows up with mathematical or algorithmic developments triggered by a specific problem, and concludes with one or two success stories, namely new insights gained thanks to these methodological developments. It also highlights some unsolved and outstanding theoretical questions, with potentially high impact on these disciplines.

Two communities will be particularly interested. The first one is the vast community of applied mathematicians and computer scientists, whose interests should be captured by the added value generated by the application of advanced concepts and algorithms to challenging biological or medical problems. The second is the equally vast community of biologists. Whether scientists or engineers, they will find in this book a clear and self-contained account of concepts and techniques from mathematics and computer science, together with success stories on their favorite systems. The variety of systems described will act as an eye opener on a panoply of complementary conceptual tools. Practically, the resources listed at the end of each chapter (databases, software) will prove invaluable to get started on a specific topic.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. Projets Exploratoires Pluridisciplinaires from CNRS/Inria/INSERM

Title: Modeling Large Protein Assemblies with Toleranced Models

Type: Projet Exploratoire Pluri-disciplinaire (PEPS) CNRS / Inria / INSERM

Duration: two years

Coordinator: F. Cazals (Inria, ABS)

Others partners: V.Doye (Inst. Jacques Monod)

Abstract: Reconstruction by Data Integration (RDI) is an emerging paradigm to reconstruct large protein assemblies, as discussed in section 5.1.3.

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.

7.2. European Initiatives

7.2.1. FP7 Projects

7.2.1.1. 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/

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.*

7.3. International Initiatives

7.3.1. Inria International Partners

7.3.1.1. Declared Inria International Partners

ABS has regular international collaboration, in particular with the members of the FP7 project *Computational geometric learning* mentioned in section 7.2.1.

7.4. International Research Visitors

7.4.1. Internships

- Angeliki Kalamara, from the University of Athens, performed a 5 month internship under the dual supervision of F. Cazals and I. Emiris (Univ. of Athens). The topic was *Modeling cryo-electron microscopy density maps*.

8. Dissemination

8.1. Scientific Animation

8.1.1. Organization of Schools and Courses

- **(Winter school Algorithms in Structural Bio-informatics)** Together with J. Cortès from LAAS / CNRS (Toulouse) and C. Robert (IBPC/CNRS), F. Cazals organized 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 model the flexibility of proteins, with a focus on methodological developments. The audience consisted of 30 PhD students and post-docs from all over the world.
- (**Statistical Learning Theory: a Short Course**) F. Cazals organized a mini-course by P. Grunwald, from the CWI. The details can be found at https://team.inria.fr/abs/statistical-learning-theory-a-short-course/.

8.1.2. Conference Program Committees

– F. Cazals was member of the following PC:

- Symposium on Geometry Processing.
- Computer Graphics International
- ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics
- International Conference on Pattern Recognition in Bioinformatics
- Computational Intelligence Methods for Bioinformatics and Biostatistics

8.1.3. Appointments

– F. Cazals was appointed as Panel expert for the 7th Framework Programme 7 from the EU, Information and Communication Technologies /

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

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

Master: F. Cazals, *Algorithmic problems in computational structural biology*, 24h, Master of Science in Computational Biology from the University of Nice Sophia Antipolis, France. (http://cbb.unice.fr))

Graduate level: F. Cazals and C. Robert, *Analyzing conformational landscapes, with applications to the design of collective coordinates*, 6h, Winter school *Algorithms in Structural Bio-informatics*.

8.2.2. Supervision

[4] http://algosb.sciencesconf.org/

(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.

(PhD thesis, ongoing) S. Marillet, *Modeling antibody - antigen complexes*, Univ. of Nice Sophia Antipolis. The thesis is co-advised by F. Cazals and P. Boudinot (INRA Jouy-en-Josas).

9. Bibliography

Major publications by the team in recent years

[1] 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, pp. 677–692

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

[3] 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

[4] 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

[5] 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

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

[7] 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, pp. 551–565, Preliminary version as Inria Tech report 6049

[8] 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, pp. 2082–2092

[9] 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, pp. 2125–2136

[10] 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, pp. 487–498

[11] N. MALOD-DOGNIN, A. BANSAL, F. CAZALS. *Characterizing the Morphology of Protein Binding Patches*, in "Proteins: structure, function, and bioinformatics", 2012, vol. 80, n° 12, pp. 2652–2665

Publications of the year

Articles in International Peer-Reviewed Journals

[12] R. CASTRO, L. JOUNEAU, H. PHAM, O. BOUCHEZ, V. GIUDICELLI, M. 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 Pathog", 2013, vol. 9, n° 1, http://hal.inria.fr/hal-00779835

[13] T. DREYFUS, V. DOYE, F. CAZALS. *Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Sub-complexes*, in "Proteins: Structure, Function, and Bioinformatics", 2013, In press [*DOI :* 10.1002/PROT.24313], http://hal.inria.fr/hal-00849795

International Conferences with Proceedings

[14] D. AGARWAL, J. ARAUJO, C. CAILLOUET, F. CAZALS, D. COUDERT, S. PERENNES. *Connectivity Inference in Mass Spectrometry based Structure Determination*, in "European Symposium on Algorithms", Sophia-Antipolis, France, France, H. BODLAENDER, G. ITALIANO (editors), Lecture Notes in Computer Science - LNCS, Springer, 2013, vol. 8125, pp. 289-300 [*DOI :* 10.1007/978-3-642-40450-4_25], http://hal.inria.fr/hal-00849873

Books or Proceedings Editing

[15] F. CAZALS, P. KORNPROBST (editors). , *Modeling in Computational Biology and Medicine: A Multidisciplinary Endeavor*, Springer, 2013, 315 p. [*DOI :* 10.1007/978-3-642-31208-3], http://hal.inria.fr/hal-00845616

Research Reports

[16] D. AGARWAL, J. ARAUJO, C. CAILLOUET, F. CAZALS, D. COUDERT, S. PÉRENNES. , *Connectivity Inference in Mass Spectrometry based Structure Determination*, Inria, June 2013, n° RR-8320, 23 p. , http://hal.inria.fr/hal-00837496

[17] D. AGARWAL, F. CAZALS, N. MALOD-DOGNIN. , *Stoichiometry Determination for Mass-spectrometry Data: the Interval Case*, Inria, February 2013, n° RR-8101, 52 p. , http://hal.inria.fr/hal-00741491

[18] F. CAZALS, T. DREYFUS, S. SACHDEVA, S. NISARG. , *Greedy Geometric Optimization Algorithms for Collection of Balls*, Inria, January 2013, n° RR-8205, 26 p. , http://hal.inria.fr/hal-00777892

[19] F. CAZALS, A. ROTH, C. ROBERT, M. CHRISTIAN. , *Towards Morse Theory for Point Cloud Data*, Inria, July 2013, n° RR-8331, 37 p. , http://hal.inria.fr/hal-00848753

Other Publications

[20] C. GARATE, S. ZAIDENBERG, J. BADIE, F. BREMOND. , *Group Tracking and Behavior Recognition in Long Video Surveillance Sequences*, January 2014, VISAPP - 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, http://hal.inria.fr/hal-00879734

References in notes

[21] 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, pp. 683-694

[22] 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° 7170, pp. 695–701

[23] 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, pp. 11.1–11.35

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

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

[26] M. L. CONNOLLY. *Analytical molecular surface calculation*, in "J. Appl. Crystallogr.", 1983, vol. 16, n° 5, pp. 548–558

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

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

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

[30] 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, pp. 531–539

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

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

[33] V. K. KRIVOV, M. KARPLUS. *Hidden complexity of free energy surfaces for peptide (protein) folding*, in "PNAS", 2004, vol. 101, n° 41, pp. 14766-14770

[34] 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

[35] 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, pp. 584–595

[36] 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, pp. 511–520

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

[38] 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, pp. 897–907

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

[40] 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, pp. 57-62 [*DOI :* 10.1073/PNAS.0407280102]

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

[42] 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, pp. 18551-18555

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

[44] 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, pp. 859-883

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

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