RESEARCH CENTRE

Nancy - Grand Est

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2020 ACTIVITY REPORT

Project-Team CAPSID

Computational Algorithms for Protein Structures and Interactions

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Contents

Project-Team CAPSID

Creation of the Team: 2015 January 01, updated into Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.1.9. – Database

A3.1.10. – Heterogeneous data

A3.1.11. – Structured data

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.2.6. – Linked data

A3.3.2. – Data mining

A3.5.1. – Analysis of large graphs

A6.1.4. – Multiscale modeling

A6.2.7. – High performance computing

A6.3.3. – Data processing

A6.5.5. – Chemistry

A8.2. – Optimization

A9.1. – Knowledge

A9.2. – Machine learning

Other research topics and application domains

B1.1.1. – Structural biology

B1.1.2. – Molecular and cellular biology

B1.1.7. – Bioinformatics

B2.2.1. – Cardiovascular and respiratory diseases

B2.2.4. – Infectious diseases, Virology

B2.4.1. – Pharmaco kinetics and dynamics

1 Team members, visitors, external collaborators

Research Scientists

- Marie-Dominique Devignes [Team leader, CNRS, Researcher, HDR]
- Isaure Chauvot de Beauchêne [CNRS, Researcher]
- Bernard Maigret [CNRS, Emeritus]

Faculty Members

- Sabeur Aridhi [Telecom Nancy, Associate Professor]
- Malika Smaïl-Tabbone [Univ de Lorraine, Associate Professor, HDR]

Post-Doctoral Fellows

- Amina Ahmed Nacer [Univ de Lorraine, until Dec 2020]
- Camille Besançon [Inria, from Mar 2020]
- Dominique Mias-Lucquin [Univ de Lorraine]

PhD Students

- Diego Amaya Ramirez [Inria]
- Adrien Blanche [Univ de Lorraine, from Nov 2020]
- Kévin Dalleau [Univ de Lorraine, from Oct 2020, Contrat ATER]
- Hrishikesh Dhondge [CNRS]
- Wissem Inoubli [Univ de Lorraine, until Aug 2020]
- Kamrul Islam [Univ de Lorraine]
- Anna Kravchenko [CNRS]
- Antoine Moniot [Univ de Lorraine]
- Bishnu Sarker [Inria]
- Athenais Vaginay [Univ de Lorraine]

Technical Staff

- Emmanuel Bresso [CNRS, Engineer, until Jul 2020]
- Kévin Dalleau [CNRS, Engineer, until Sep 2020]
- Philippe Noel [Inria, Engineer, until Feb 2020]
- Louane Sigrist [Univ de Lorraine, Engineer, from Dec 2020]

Interns and Apprentices

- Salma Aziz [Inria, from Apr 2020 until Jul 2020]
- Alix Delannoy [Univ de Lorraine, from Oct 2020]
- Valentin Retter [Univ de Lorraine, from Apr 2020 until Jun 2020]
- Karina Mayumi Sakita [CNRS, until Sep 2020]
- Amal Stiti [Univ de Lorraine, from Oct 2020]

Administrative Assistants

- Antoinette Courrier [CNRS]
- Isabelle Herlich [Inria]

External Collaborators

- Taha Boukhobza [Univ de Lorraine]
- Sjoerd Jacob De Vries [INSERM]

2 Overall objectives

2.1 Computational Challenges in Structural Biology

Many of the processes within living organisms can be studied and understood in terms of biochemical interactions between large macromolecules such as DNA, RNA, and proteins. To a first approximation, DNA may be considered to encode the blueprint for life, whereas proteins and RNA make up the three-dimensional (3D) molecular machinery. Many biological processes are governed by complex systems of proteins and/or RNA which interact cooperatively to regulate the chemical composition within a cell or to carry out a wide range of biochemical processes such as photosynthesis, metabolism, and cell signalling, for example. It is becoming increasingly feasible to isolate and characterise some of the individual molecular components of such systems, but it still remains extremely difficult to achieve detailed models of how these complex systems actually work. Consequently, a new multidisciplinary approach called integrative structural biology has emerged which aims to bring together experimental data from a wide range of sources and resolution scales in order to meet this challenge [73, 58].

Understanding how biological systems work at the level of 3D molecular structures presents fascinating challenges for biologists and computer scientists alike. Despite being made from a small set of simple chemical building blocks, protein and nucleic acid (NA) molecules have a remarkable ability to self-assemble into complex molecular machines which carry out very specific biological processes. As such, these molecular machines may be considered as complex systems because their properties are much greater than the sum of the properties of their component parts.

2.2 Two Research Axes

The overall objective of the Capsid team is to develop algorithms and software to help study biological systems and phenomena from a structural point of view. In particular, the team aims to develop algorithms which can help to model the structures of large multi-component biomolecular machines and to develop tools and techniques to represent and mine knowledge of the 3D shapes of proteins, NA and their interactions. Thus, a unifying theme of the team is to tackle the recurring problem of representing and reasoning about large 3D macromolecular shapes. More specifically, our aim is to develop computational techniques to represent, analyse, and compare the shapes and interactions of biomolecules in order to help better understand how their 3D structures relate to their biological function. In summary, the Capsid team is organised according to two research axes whose complementarity constitutes an original contribution to the field of structural bioinformatics:

- Axis 1: New Approaches for Knowledge Discovery in Structural Databases,
- Axis 2: Integrative Multi-Component Assembly and Modeling.

As indicated above, structural biology is largely concerned with determining the 3D atomic structures of proteins and NA molecules, and then using these structures to study their biological properties and interactions. Each of these activities can be extremely time-consuming. Solving the 3D structure of even a single protein using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy can often take many months or even years of effort. Even simulating the interaction between two proteins using a detailed atomistic molecular dynamics simulation can consume many thousands of CPU-hours. While most X-ray crystallographers, NMR spectroscopists, and molecular modelers often use conventional sequence and structure alignment tools to help propose initial structural models through the homology principle, they often study only individual structures or interactions at a time. Due to the difficulties outlined above, only relatively few research groups are able to solve the structures of large multi-component systems.

Similarly, most current algorithms for comparing protein structures, and especially those for modeling protein interactions, work only at the pair-wise level. Of course, such calculations may be accelerated considerably by using dynamic programming (DP) or fast Fourier transform (FFT) techniques. However, it remains extremely challenging to scale up these techniques to model multi-component systems. For example, the use of high performance computing (HPC) facilities may be used to accelerate arithmetically intensive shape-matching calculations, but this generally does not help solve the fundamentally combinatorial nature of many multi-component problems. It is therefore necessary to devise heuristic hybrid approaches which can be tailored to exploit various sources of domain knowledge. We therefore set ourselves the following main computational objectives:

- classify and mine protein structures and interactions,
- develop multi-component assembly techniques for integrative structural biology.

3 Research program

3.1 Classifying and Mining Protein Structures and Protein Interactions

3.1.1 Context

The scientific discovery process is very often based on cycles of measurement, classification, and generalisation. It is easy to argue that this is especially true in the biological sciences. The proteins that exist today represent the molecular product of some three billion years of evolution. Therefore, comparing protein sequences and structures is important for understanding their functional and evolutionary relationships [69, 50]. There is now overwhelming evidence that all living organisms and many biological processes share a common ancestry in the tree of life. Historically, much of bioinformatics research has focused on developing mathematical and statistical algorithms to process, analyse, annotate, and compare protein and DNA sequences because such sequences represent the primary form of information in biological systems. However, there is growing evidence that structure-based methods can help to predict networks of protein-protein interactions (PPIs) with greater accuracy than those which do not use structural evidence [54, 74]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [41].

3.1.2 Formalising and Exploiting Domain Knowledge

Concerning protein structure classification, we aim to explore novel classification paradigms to circumvent the problems encountered with existing hierarchical classifications of protein folds and domains. In particular it will be interesting to set up fuzzy clustering methods taking advantage of our previous work on gene functional classification [45], but instead using Kpax domain-domain similarity matrices. A non-trivial issue with fuzzy clustering is how to handle similarity rather than mathematical distance matrices, and how to find the optimal number of clusters, especially when using a non-Euclidean similarity measure. We will adapt the algorithms and the calculation of quality indices to the Kpax similarity

measure. More fundamentally, it will be necessary to integrate this classification step in the more general process leading from data to knowledge called Knowledge Discovery in Databases (KDD) [48].

Another example where domain knowledge can be useful is during result interpretation: several sources of knowledge have to be used to explicitly characterise each cluster and to help decide its validity. Thus, it will be useful to be able to express data models, patterns, and rules in a common formalism using a defined vocabulary for concepts and relationships. Existing approaches such as the Molecular Interaction (MI) format [51] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [60]. A different point of view is represented in the Interaction Network Ontology (INO), a community-driven ontology that aims to standardise and integrate data on interaction networks and to support computer-assisted reasoning [61]. However, this ontology does not integrate basic 3D concepts and structural relationships. Therefore, extending such formalisms and symbolic relationships will be beneficial, if not essential, when classifying the 3D shapes of proteins at the domain family level.

Domain family classification is also relevant for studying domain-domain interactions (DDI). Our previous work on Knowledge-Based Docking (KBDOCK, [6, 4]) will be updated and extended using newly published DDIs. Methods for inferring new DDIs from existing protein-protein interactions (PPIs) will be developed. Efforts should be made for validating such inferred DDIs so that they can be used to enrich DDI classification and predict new PPIs.

In parallel, we also intend to design algorithms for leveraging information embedded in biological knowledge graphs (also known as complex networks). Knowledge graphs mostly represent PPIs, integrated with various properties attached to proteins, such as pathways, drug binding or relation with diseases. Setting up similarity measures for proteins in a knowledge graph is a difficult challenge. Our objective is to extract useful knowledge from such graphs in order to better understand and highlight the role of multi-component assemblies in various types of cell or organisms. Ultimately, knowledge graphs can be used to model and simulate the functioning of such molecular machinery in the context of the living cell, under physiological or pathological conditions.

3.1.3 Function Annotation in Large Protein Graphs

Knowledge of the functional properties of proteins can shed considerable light on how they might interact. However, huge numbers of protein sequences in public databases such as UniProt/TrEMBL lack any functional annotation, and the functional annotation of such sequences is a highly challenging problem. We are developing graph-based and machine learning techniques to annotate automatically the available unannotated sequences with functional properties such as EC numbers and Gene Ontology (GO) terms (note that these terms are organised hierarchically allowing generalization/specialization reasoning). The idea is to transfer annotations from expert-reviewed sequences present in the UniProt/SwissProt database (about 560 thousands entries) to unreviewed sequences present in the UniProt/TrEMBL database (about 80% of 180 millions entries). For this, we have to learn from the UniProt/SwissProt database how to compute the similarity of proteins sharing identical or similar functional annotations. Various similarity measures can be tested using cross-validation approches in the UniProt/SwissProt database. For instance, we can use primary sequence or domain signature similarities. More complex similarities can be computed with graph-embedding techniques.

This work is in progress with Bishnu Sarker's PhD project and a first approach called GrAPFI (Graph-based Automatic Protein Function Inference) was presented at conferences in 2018 [14, 16].

3.2 Integrative Multi-Component Assembly and Modelling

3.2.1 Context

At the molecular level, each biomolecular interaction is embodied by a physical 3D interface. Therefore, if the 3D structures of a pair of interacting protein/NA molecules are known, it should in principle be possible for a docking algorithm to use this knowledge to predict the structure of the complex. However, modeling protein and even more RNA flexibility accurately during docking is very computationally expensive. This is due to the very large number of internal degrees of freedom in each molecule, associated with twisting motions around covalent bonds. Therefore, it is highly impractical to use detailed force-field or geometric representations in a brute-force docking search. Instead, most docking algorithms use fast

heuristic methods to perform an initial rigid-body search in order to locate a relatively small number of candidate binding orientations, and these are then refined using a more expensive interaction potential or force-field model, which might also include flexible refinement using molecular dynamics (MD), for example.

3.2.2 Polar Fourier Docking Correlations

In our *Hex* protein docking program [64], the shape of a protein molecule is represented using polar Fourier series expansions of the form

$$\sigma(\underline{x}) = \sum_{nlm} a_{nlm} R_{nl}(r) y_{lm}(\theta,\phi), \tag{1}$$

where $\sigma(\underline{x})$ is a 3D shape-density function, a_{nlm} are the expansion coefficients, $R_{nl}(r)$ are orthonormal Gauss-Laguerre polynomials and $y_{lm}(\theta,\phi)$ are the real spherical harmonics. The electrostatic potential, $\phi(\underline{x})$, and charge density, $\rho(\underline{x})$, of a protein may be represented using similar expansions. Such representations allow the *in vacuo* electrostatic interaction energy between two proteins, A and B, to be calculated as [53]

$$E = \frac{1}{2}\int \phi_A(\underline{x})\rho_B(\underline{x})\mathrm{d}\underline{x} + \frac{1}{2}\int \phi_B(\underline{x})\rho_A(\underline{x})\mathrm{d}\underline{x}. \tag{2}$$

This equation can be demonstrated using the notion of *overlap* between 3D scalar quantities to give a physics-based scoring function. If the aim is to find the configuration that gives the most favourable interaction energy, then it is necessary to perform a six-dimensional search in the space of available rotational and translational degrees of freedom. By re-writing the polar Fourier expansions using complex spherical harmonics, we showed previously that fast Fourier transform (FFT) techniques may be used to accelerate the search in up to five of the six degrees of freedom [65]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [10, 8]. Consequently, we are continuing to explore new ways to exploit the polar Fourier approach.

3.2.3 Assembling Symmetrical Protein Complexes

Although protein-protein docking algorithms are improving [63, 55], it still remains challenging to produce a high resolution 3D model of a protein complex using *ab initio* techniques. This is mainly due to the problem of structural flexibility described above. However, with the aid of even just one simple constraint on the docking search space, the quality of docking predictions can improve considerably [10], [65]. In particular, many protein complexes involve symmetric arrangements of one or more subunits, and the presence of symmetry may be exploited to reduce the search space considerably [40, 62, 68]. For example, using our operator notation (in which $\hat{R}$ and $\hat{T}$ represent 3D rotation and translation operators, respectively), we have developed an algorithm which can generate and score candidate docking orientations for monomers that assemble into cyclic (C_n) multimers using 3D integrals of the form

$$E_{AB}(y,\alpha,\beta,\gamma) = \int \left[\hat{T}(0,y,0)\hat{R}(\alpha,\beta,\gamma)\phi_A(\underline{x})\right] \times \left[\hat{R}(0,0,\omega_n)\hat{T}(0,y,0)\hat{R}(\alpha,\beta,\gamma)\rho_B(\underline{x})\right]\mathrm{d}\underline{x}, \tag{3}$$

where the identical monomers A and B are initially placed at the origin, and $\omega_n = 2\pi/n$ is the rotation about the principal n-fold symmetry axis. This example shows that complexes with cyclic symmetry have just 4 rigid body degrees of freedom (DOFs), compared to $6(n-1)$ DOFs for non-symmetrical n-mers. We have generalised these ideas in order to model protein complexes that crystallise into any of the naturally occurring point group symmetries (C_n, D_n, T, O, I). This approach was published in 2016 [11], and was subsequently applied to several symmetrical complexes from the "CAPRI" blind docking experiment [47]. Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to build and refine candidate solutions using a more accurate coarse-grained (CG) force-field scoring function.

3.2.4 Coarse-Grained Models

Many approaches have been proposed in the literature to take into account protein (and more recently RNA/DNA) flexibility during docking. The most thorough methods rely on expensive atomistic simulations using MD. However, much of a MD trajectory is unlikely to be relevant to a docking encounter

unless it is constrained to explore a putative protein-protein/NA interface. Consequently, MD is normally only used to refine a small number of candidate rigid body docking poses. A much faster, but more approximate method is to use "coarse-grained" (CG) normal mode analysis (NMA) techniques to reduce the number of flexible degrees of freedom to just one or a handful of the most significant vibrational modes [59, 46, 56, 57]. In our experience, docking ensembles of NMA conformations do not give much improvement over basic FFT-based soft docking [72], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [5].

In the last few years, CG force-field models have become increasingly popular in the MD community because they allow very large biomolecular systems to be simulated using conventional MD programs [39]. Typically, a CG force-field representation replaces the atoms in each amino acid with from 2 to 4 “pseudo-atoms”, and it assigns each pseudo-atom a small number of parameters to represent its chemo-physical properties. By directly attacking the quadratic nature of pair-wise energy functions, coarse-graining can speed up MD simulations by up to three orders of magnitude. Nonetheless, such CG models can still produce useful models of very large multi-component assemblies [67]. Furthermore, this kind of CG model effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [52]. We are currently developing a CG scoring function for RNA-protein docking by fragments assembly. This work is part of the PhD project of Anna Kravchenko.

3.2.5 Assembling Multi-Component Complexes and Integrative Structure Modelling

We also want to develop related approaches for integrative structure modeling using cryo-electron microscopy (cryo-EM). Thanks to recent developments in cryo-EM instruments and technologies, it is now feasible to capture low resolution images of very large macromolecular machines. However, while such developments offer the intriguing prospect of being able to trap biological systems in unprecedented levels of detail, there will also come with an increasing need to analyse, annotate, and interpret the enormous volumes of data that will soon flow from the latest instruments. In particular, a new challenge that is emerging is how to fit previously solved high resolution protein structures into low resolution cryo-EM density maps. However, the problem here is that large molecular machines will have multiple sub-components, some of which will be unknown, and many of which will fit each part of the map almost equally well. Thus, the general problem of building high resolution 3D models from cryo-EM data is like building a complex 3D jigsaw puzzle in which several pieces may be unknown or missing, and none of which will fit perfectly. We wish to proceed firstly by putting more emphasis on the single-body terms in the scoring function [44], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space. This work has made much progress during the PhD project of Maria Elisa Ruiz Echartea but still requires further efforts.

3.2.6 Protein-Nucleic Acid Interactions

As well as playing an essential role in the translation of DNA into proteins, RNA molecules carry out many other essential biological functions in cells, often through their interactions with proteins. A critical challenge in modeling such interactions computationally is that the RNA is often highly flexible, especially in single-stranded (ssRNA) regions of its structure. These flexible regions are often very important because it is through their flexibility that the RNA can adjust its 3D conformation in order to bind to a protein surface. However, conventional protein-protein docking algorithms generally assume that the 3D structures to be docked are rigid, and so are not suitable for modeling protein-RNA interactions. There is therefore much interest in developing dedicated protein-RNA docking algorithms which can take RNA flexibility into account. This research topic has been initiated with the recruitement of Isaure Chauvot de Beauchêne in 2016 and is becoming a major activity in the team. A novel flexible docking algorithm is currently under development in the team. It first docks small fragments of ssRNA (typically three nucleotides at a time) onto a protein surface, and then combinatorially reassembles those fragments in order to recover a contiguous ssRNA structure on the protein surface [43, 42].

As the correctness of the initial docking of the fragments settles an upper limit to the correctness of the full model, we are now focusing on improving that step. A key component of our docking tool is the energy function of the protein-fragment interactions that is used both to drive the sampling (positioning of the fragments) by minimization, and to discriminate the correct final positions from decoys (i.e., false

positives). We are developing a new approach to create knowledge-based parameters for coarse-grain energy functions from public structural data, in collaboration with Sjoerd de Vries (INSERM). Such approach will be applied first to ssRNA-protein complexes, then to other types of complexes such as protein-peptides.

Another key requirement for this approach is an exhaustive but non-redundant library of possible internal conformations of RNA fragments. Our library is built by clustering hundreds of thousands of experimentally known RNA structures, based on an approximate geometric similarity criteria. We are currently developping new exact approaches for the clustering of 3D conformations based on internal coordinates, in order to optimise the representativity of the library. This is part of the PhD subject of Antoine Moniot, co-supervised by Yann Guermeur (ABC LORIA team).

In the future, we will improve the combinatorial algorithm used for reassembling the docked fragments using both experimental constraints (PhD project Anna Kravchenko) and machine-learning approaches (PhD project Hrishikesh Dondge).

4 Application domains

4.1 Biomedical Knowledge Discovery

Participants Marie-Dominique Devignes *(contact person)*, Malika Smaïl-Tabbone *(contact person)*, Sabeur Aridhi, David Ritchie, Gabin Personeni, Seyed Ziaeddin Alborzi, Kevin Dalleau, Bishnu Sarker, Emmanuel Bresso, Claire Lacomblez, Floriane Odje, Athénaïs Vaginay.

Our main application for Axis 1 : "New Approaches for Knowledge Discovery in Structural Databases", concerns biomedical knowledge discovery. We intend to develop KDD approaches on preclinical (experimental) or clinical datasets integrated with knowledge graphs with a focus on discovering which PPIs or molecular machines play an essentiel role in the onset of a disease and/or for personalised medicine.

As a first step we have been involved since 2015 in the ANR RHU “FIGHT-HF” (Fight Heart Failure) project, which is coordinated by the CIC-P (Centre d’Investigation Clinique Plurithématique) at the CHRU Nancy and INSERM U1116. In this project, the molecular mechanisms that underly heart failure (HF) are re-visited at the cellular and tissue levels in order to adapt treatments to patients’ needs in a more personalised way. The Capsid team is in charge of a workpackage dedicated to network science. A platform has been constructed with the help of a company called Edgeleap (Utrecht, NL) in which biological molecular data and ontologies, available from public sources, are represented in a single integrated complex network also known as knowledge graph. We are developing querying and analysis facilities to help biologists and clinicians interpreting their cohort results in the light of existing interactions and knowledge. We are also currently analysing pre-clinical data produced at the INSERM unit on the comparison of aging process in obese versus lean rats. Using our expertise in receptor-ligand docking, we are investigating possible cross-talks between mineralocorticoid and other nuclear receptors.

Another application is carried out in the context of an interdisciplinary project funded by the Université de Lorraine, in collaboration with the CRAN laboratory. It concerns the study of the role of estrogen receptors in the development of gliobastoma tumors. The available data is high-dimensional but involves rather small numbers of samples. The challenge is to identify relevant sets of genes which are differentially expressed in various phenotyped groups (w.r.t. gender, age, tumor grade). The objectives are to infer pathways involving these genes and to propose candidate models of tumor development which will be experimentally tested thanks to an ex-vivo experimental system available at the CRAN.

Finally, simulating biological networks will be important to understand biological systems and test new hypotheses. One major challenge is the identification of perturbations responsible for the transformation of a healthy system to a pathological one and the discovery of therapeutic targets to reverse this transformation. Control theory, which consists in finding interventions on a system in order to prevent it to go in undesirable states or to force it to converge towards a desired state, is of great interest for this challenge. It can be formulated as “How to force a broken system (pathological) to act as it should do (normal state)?”. Many formalisms are used to model biological processes, such as Differential Equations

(DE), Boolean Networks (BN), cellular automata. In her PhD thesis, Athenaïs Vaginay investigates ways to find a BN fitting both the knowledge about topology and state transitions “inferred“ from experimental data. This step is known as “boolean function synthesis”. Our aim is to design automated methods for building biological networks and define operators to intervene on them [71]. Our approaches will be driven by knowledge and will keep close connection with experimental data.

4.2 Prokaryotic Type IV Secretion Systems

Participants Marie-Dominique Devignes *(contact person)*, Isaure Chauvot de Beauchêne *(contact person)*, Bernard Maigret, Philippe Noel, Dominique Mias-Lucquin.

Concerning Axis 2 : "Integrative Multi-Component Assembly and Modeling", our first application domain is related to prokaryotic Type IV secretion systems.

Prokaryotic type IV secretion systems constitute a fascinating example of a family of nanomachines capable of translocating DNA and protein molecules through the cell membrane from one cell to another [38]. The complete system involves at least 12 proteins. The structure of the core channel involving three of these proteins has recently been determined by cryo-EM experiments for Gram-negative bacteria [49, 66]. However, the detailed nature of the interactions between the other components and the core channel remains to be found. Therefore, these secretion systems represent a family of complex biological systems that call for integrated modeling approaches to fully understand their machinery.

In the framework of the Lorraine Université d’Excellence (LUE-FEDER) “CITRAM” project we are pursuing our collaboration with Nathalie Leblond of the Genome Dynamics and Microbial Adaptation (DynAMic) laboratory (UMR 1128, Université de Lorraine, INRAE) on the mechanism of horizontal transfer by integrative conjugative elements (ICEs) and integrative mobilisable elements (IMEs) in prokaryotic genomes. These elements use type IV secretion systems for transferring DNA horizontally from one cell to another. We have discovered more than 200 new ICEs/IMEs by systematic exploration of 72 Streptococcus genomes and characterised a new class of relaxases [70]. We have modeled the dimer of this relaxase protein by homology with a known structure. For this, we have created a new pipeline to model symmetrical dimers of multi-domains proteins. As one activity of the relaxase is to cut the DNA for its transfer, we are also currently studying the DNA-protein interactions that are involved in this very first step of horizontal transfer (see next section).

4.3 Protein - RNA Interactions

Participants Isaure Chauvot de Beauchêne *(contact person)*, Antoine Moniot, Anna Kravchenko, Hrishikesh Dhondge, Alix Delannoy, Marie-Dominique Devignes, Malika Smaïl-Tabbone.

The second application domain of Axis 2 concerns protein-nucleic acid interactions. We need to assess and optimise our new algorithms on concrete protein-nucleic acid complexes in close collaboration with external partners coming from the experimental field of structural biology. To facilitate such collaborations, we are creating automated and re-usable protein-nucleic acid docking pipelines.

This is the case for our PEPS collaboration “InterANRIL” with the IMoPA lab (CNRS-Université de Lorraine). We are currently working with biologists to apply our fragment-based docking approach to model complexes of the long non-coding RNA (lncRNA) ANRIL with proteins and DNA.

In the framework of our LUE-FEDER CITRAM project (see above), we are adapting this approach and pipeline to single-strand DNA docking, in order to model the complex formed by a bacterial relaxase and its target DNA.

In the framework of our H2020 ITN project RNAct, we tackle a defined group of RNA-binding proteins containing RNA-Recognition Motifs (RRM). We study existing and predicted complexes between various types of RRMs and various RNA sequences in order to infer rules of their sequence-structure-interaction

relationship, and to help design new synthetic proteins with targeted RNA specificity. This work is made in tight collaboration with computer scientists and biophysicists of the consortium.

5 Social and environmental responsibility

5.1 Environmental Footprint of Research Activities

Many conferences have been run online in 2020. This has lead to a strong decrease in the environmental footprint of the team with respect to plane travels. In particular four trips to Spain for the ECCB 2020 (in September) and the IWBBIO 2020 (in October) conferences have been cancelled.

6 Highlights of the year

6.1 Anti-Covid Research

The team has designed and carried out two projects related to the fight against the Covid-19 pandemia.

- Virtual screening and molecular dynamics to find inhibitors of SARS-COV-2 virus binding to its receptor
- Docking and 3D modeling of the binding of the SARS-COV-2 nsp9 protein to ssRNA

Unfortunately, neither of these projects was selected and therefore supported by the Mission Inria Covid-19. No report or publication is available so far.

7 New software and platforms

7.1 New software

7.1.1 ProtNAff

Name: Protein - Nucleic Acids Filters and Fragments

Keywords: Structural alphabet, Structural Biology, Nucleic Acids

Scientific Description: The modeling of nucleic acids (NA) - protein interactions can greatly help the design of therapeutic NA. Atomistic models of NA fragments can be used to model the 3D structures of NA-protein complexes, to subdivide the handling of RNAs great flexibility. One way to obtain relevant RNA fragments is to extract them from existing 3D structures of interactions corresponding to the context one wants to model (surface area surrounding NA 2D structures, specific protein families, specific sequences) and to learn from them.

Functional Description: ProtNAff is a python-based software for (i) the automated parsing, correction and annotation of all protein-nucleic acid structures in the public Protein Data Bank, (ii) the creation of libraries of non-redundant RNA/DNA structural fragments, (iii) the selection of sets of structures by customised queries, and (iv) the computation of statistics on sets of RNA/DNA - protein structures.

URL: https://github.com/isaureCdB/NAfragDB

Publication: hal-02393039

Contacts: Isaure Chauvot de Beauchêne, Antoine Moniot, Sjored Jacob De Vries

Participants: Isaure Chauvot de Beauchêne, Antoine Moniot, Sjored Jacob De Vries

Partner: INSERM

7.1.2 InteR3Mdb

Name: Database for interactions between RNA and RRM (RNA Recognition Motif)

Keywords: Databases, Proteins, Nucleic Acids, 3D interaction, Biological sequences

Functional Description: InteR3Mdb is a comprehensive database gathering and interconnecting public data on the 3D structures and sequences of a very conserved protein domain, the RNA Recognition Motif (RRM), and its interaction with RNA. It comprises a web-interface, a query interface and an API interface.

Release Contributions: This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 813239. It will be confidential, for usage within the H2020 ITN project called RNAct, for the whole duration of the project (2019-2023). After the project, it will be made publicly available.

URL: https://inter3mdb.loria.fr

Contact: Hrishikesh Dhondge

7.1.3 PPIDM

Name: Protein-protein interactions (PPI) domain miner

Keywords: Proteins, 3D interaction, Protein domain, Link prediction, Graph

Functional Description: PPIDM exploits a formalization of a tripartite graph between a source of PPIs (protein-protein interactions) and two sets of domains present in the proteins involved in these PPIs. The software performs link inference between the two sets of domains by calculating a probability score for these links. A threshold is determined by learning from positive and negative examples. The DDIs (domain-domain interactions) identified by PPIDM are made available as a pool of plausible DDIs that can be the subject of subsequent analyses.

URL: https://ppidm.loria.fr

Contact: Marie-Dominique Devignes

7.2 New platforms

The CAPSID team has contributed to the outcome of the MBI-DS4H research platform for sharing resources in structural bioinformatics (MBI: Modeling Biomolecules and their Interactions) and data science for health (DS4H). This platform is shared with the ORPAILLEUR team and open for any other team working in the field. High-performance computing is possible on clusters equipped with CPU and GPU and Infiniband connectivity, especially for molecular dynamic runs. Access to Grid5K is facilitated. Storage servers have been installed. A DMZ server is available for external access to resources having reached the production status. The technical support is ensured by the LORIA SISR (Service d'Ingénierie en Soutien de la Recherche).
https://mbi.loria.fr

8 New results

8.1 Axis 1 : New Approaches for Knowledge Discovery in Structural Databases

8.1.1 Biomedical Knowledge Discovery

In collaboration with clinicians at the CHRU Nancy in the framework of the RHU FIGHT-HF program and of the Contrat d'Interface, we have developed the GKBox (Graph Knowledge Box) which integrates, in the Neo4J graph format, a huge amount of data from over 20 public biological databases such as UniProt, DisGeNET or String. In parallel we have developed automated querying mechanisms to retrieve

constrained paths from the GKBox, such paths restricted to protein-protein interactions or paths following a protein-pathway-protein pattern, and involving only the proteins in a list of interest. Such functionality has been used to better understand the mechanisms underpinning lists of proteins associated with particular phenotypes in the cohorts, as in [21, 22, 28].

Linked RDF data constitute another type of knowledge graphs that we addressed in collaboration with the Orpailleur Team in the frame of the PraktikPharma ANR project coordinated by Adrien Coulet. Our participation led to two co-signed publications: on a manually annotated corpus for pharmacogenomics on the one hand, and on a scalable method for mining path patterns from knowledge graphs on the other hand [24, 32]. A third publication was submitted at the end of 2020 on investigating adverse drug reaction mechanisms with knowledge graph mining. The PhD project of Kamrul Islam addresses the link prediction problem in knowledge graphs using deep learning approaches while taking into consideration domain knowledge and integrity constraints.

Finally we also participated in a preliminary study carried out in the Orpailleur team to explain multicriteria decision making with formal concept analysis [29], leveraging concept lattices as yet another type of knowledge graph.

8.1.2 Stochastic Random Trees for Similarity Computation

We proposed a method to compute similarities on unlabeled data, based on extremely randomised trees called Unsupervised Extremely Randomised Trees (UET) [20]. The main idea is to randomly split the data in an iterative fashion until a stopping criterion is met, and subsequently compute similarity values based on the co-occurrence of samples in the leaves of each generated tree. We evaluated our method on synthetic and real-world datasets using metrics similar to intracluster and intercluster similarities. Such metrics allow us to assess the computed similarities instead of a clustering algorithm's results. We also assessed some interesting properties of UET such as invariance under monotone transformations of variables and robustness to correlated variables and noise. Finally, we performed hierarchical agglomerative clustering on synthetic and real-world homogeneous and heterogeneous datasets using UET versus standard similarity measures. The experiments showed that the algorithm outperforms existing methods in several cases, and reduces the amount of preprocessing needed with many real-world datasets.

As an extension of the UET approach, we wanted to address the current challenge of clustering graph vertices especially in complex networks, i.e., graphs with attributed vertices or edges [30]. Thus we proposed GraphTrees, a novel method that relies on random decision trees to compute pairwise dissimilarities between vertices in a graph. We show that combining different types of trees, it is possible to extend this framework to graphs where the vertices have attributes. GraphTrees can handle heterogeneous types of node attributes as well. Here again, unlike other approaches, the attributes do not need to be preprocessed. We also show that our approach is competitive with state of the art methods in the case of non-attributed graphs in terms of quality of clustering, and provides promising results in the case of vertex-attributed graphs. By extending the use of an already well established approach – the random trees – to graphs, our approach opens interesting research directions.

8.1.3 Graph-based Approaches for Machine Learning and Protein Annotation

One of the methods developed by Bishnu Sarker for protein function annotation is based on label propagation in graphs. GrAPFI (Graph-based Automatic Protein Function annotation) relies on a protein graph in which edges are weighted by the domain similarity between proteins [26]. When applied to function annotations taken from Gene Ontology, the method was improved by post-processing of the predicted annotations, leveraging the hierarchical structure of the ontology. Moreover it was shown that the post-processing step also improves other methods of protein function annotation [33].

Collaborative work on the development of a distributed algorithm for large-scale graph clustering has led to a technical report in 2020 [36] which contributed to the completion of Wissem Inoubli's PhD thesis (defended early January 2021).

Our work on large-scale distributed graphs led us to participate in the coordination of a special issue on "Advances on large evolving graphs" in Future Generation Computer Systems [17].

8.2 Axis 2 : Integrative Multi-Component Assembly and Modeling

8.2.1 Docking of ssRNA on Proteins

An RNA hairpin is a particular RNA structure made of a double-strand helix closed on one side by a single-strand (ss) loop. We have adapted our fragment-based docking method to model the interactions of an RNA hairpin of known 2D structure and sequence to a protein of known 3D structure. Compared to the original method for linear single-stranded RNA (ssRNA), the new version integrates the constraint of loop closure in the combinatorial assembly of the ssRNA fragments docked on the protein surface. A proof of principle of this new method on a known hairpin-protein complex has been presented in a talk by Antoine Moniot (PhD) at the JOBIM2020 conference [31].

We also developed a new version of NAfragDB, renamed ProtNAff (Protein-Nucleic Acid filters and fragments), for the parsing and structural anotation of protein-NA 3D structures from the public Protein Data Bank (PDB), the selection of sets of structures based on multiple criteria (sequence, structure, type of interactions), the creation of libraries of fragment conformations, and the statistical analysis of the fragments properties. With this tool, we have created new specific benchmarks for testing hairpin-protein docking, ssRNA-protein docking and ssDNA-protein docking. We also showed that around half of the local RNA conformations in the PDB are induced by the RNA binding to the protein. ProtNAff is available on https://github.com/isaureCdB/ProtNAff.

As fragment libraries are a key component of our fragment-based docking approach, we are currently developing new clustering criteria to create libraries as representative as possible of the diversity of local RNA/DNA conformations. A new representation of RNA in internal coordinates, developed by Alix Delannoy (3A Mines research training), is showing very promising results.

In the frame of our H2020 ITN project RNAct, which aims at designing new RNA-binding proteins based on the well-conserved protein domain called RNA Recognition Motif (RRM), we created InteR3Mdb (see section 7). With this tool, we created the first general alignment of all sequences of RRMs that have a 3D structure available. These results will serve as the basis to apply ML methods on the prediction of RRMs structure from sequence and of RRM-RNA interaction from structure. We also created a pipeline for the homology modeling of RRM structure from sequence, which will be enriched by the outcome of the ML application on InteR3Mdb. All this is part of the PhD project of Hrishikesh Dhondge.

Finally, in collaboration with Sjoerd de Vries (INSERM), we have created a largely automated reactive pipeline for the modeling of protein-ssRNA complexes, starting from RNA sequence, protein structure and experimental identification of some protein residues at the binding interface. This has been applied to the study of the SARS-COV-2 nsp9 protein in the frame of a collaborative biohackathon involving students from the CAPSID team and from the RNAct project. The pipeline is available on https://github.com/sjdv1982/biohackathon-covid.

8.2.2 3D Modeling and Virtual Screening

The docking method named EROS-Dock, developed by Maria Elisa Ruiz-Echartea during her PhD thesis, has been improved by the use of restraints that lead to higher quality models in pairwise and multi-component protein docking [25].

Virtual screening of small molecules libraries is an essential part of the team's expertise that is often called upon by external partners in biology. Success has been obtained in the discovery of a new type of inhibitor of the MET kinase domain [18], and in the understanding of the catalytic mechanism and inhibitor binding of aminopeptidase A [19]. This approach is also currently experimented in the search for inhibitors of a galactosyltransferase involved in mucopolysaccharidoses [37]. Reciprocally, in silico methods have been used in a collaborative work with a Brazilian team to search for new protein targets in the fight against a plant disease affecting coffee [34].

The use of chemoinformatics methods to help antibacterial discovery has been reviewed in an Inria technical report [35].

Finally, a great deal of efforts has gone in 2020 into finding Covid-19 inibitors, in the form of either peptides mimicking the binding receptor of the virus spike protein, or chemical molecules meeting a certain number of expert criteria and capable of interfering with the binding of the spike to its receptor. Collaborations have been established with a mathematics team at the Université Savoie Mont-Blanc for analysing amino acid contacts during molecular dynamic trajectories, and with a biology team at

the CNRS-Université de Montpellier for performing biological assays to test all the candidate inhibitors. Some promising results now await full biological confirmation.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Participation in other international programs

Project Tempographs: Analysing big data with temporal graphs and machine learning. Application to urban traffic analysis and protein function annotation.

Participants Sabeur Aridhi *(contact person)*, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Bishnu Sarker, Wissem Inoubli.

Partners: LORIA/Inria NGE, Federal University of Ceará (UFC).
Value: 20 k€.
Duration: 2017–2020.
Description: This project aims to investigate and propose solutions for both urban traffic-related problems and protein annotation problems. In the case of urban traffic analysis, problems such as traffic speed prediction, travel time prediction, traffic congestion identification and nearest neighbors identification will be tackled. In the case of protein annotation problem, protein graphs and/or protein–protein interaction (PPI) networks will be modeled using dynamic time-dependent graph representations.

9.2 European initiatives

9.2.1 FP7 & H2020 Projects

H2020 ITN RNAct

Participants Isaure Chauvot de Beauchêne *(contact person)*, Marie-Dominique Devignes, Malika Smail-Tabbone, Anna Kravchenko, Hrishikesh Dhondge.

Program: H2020 Innovative Training Network
Project acronym:RNActProject
Title: Enabling proteins with RNA recognition motifs for synthetic biology and bio-analytics
Duration: October 2018 - October 2022
Coordinator: Wim Vranken (Vrije University Bruxelles, Belgium)
Other partners:

- LORIA, CNRS (France),
- Helmholtz Center Munich (Germany),
- Conseio Superior de Investigaciones Cientificas, Instituto de Biologia Molecular y Celular de Plantas (Spain),
- Ridgeview instruments AB (Sweden),
- Giotto Biotech Srl (Italy),
- Dynamic Biosensors GmbH (Germany).

Abstract: This project aims at designing new proteins with "RNA recognition motifs (RRM)" that target a specific RNA, for exploitation in synthetic biology and bio-analytics. It combines approaches from sequence-based and structure-based computational biology with experimental biophysics, molecular biology and systemic biology. Our scientific participation regards the creation and usage of a large database on RRMs for KDD, and the development of RNA-protein docking methods.
URL: http://rnact.eu

9.2.2 Collaborations with major European organizations

ELIXIR: 3D-bioinfo Community. We are members of the new ELIXIR 3D-bioinfo community. More specifically, Isaure Chauvot de Beauchêne is member of the sub-section known as *Tools to describe, analyse, annotate, and predict nucleic acid structures* of this community. ELIXIR Communities enable the participation of communities of practice in different areas of the life sciences in the activities of ELIXIR. The goal is to underpin the evolution of data, tools, interoperability, compute and training infrastructures for European life science informatics (see https://www.elixir-europe.org/use-cases).

ELIXIR: Interoperability Platform Marie-Dominique Devignes is collaborating with the ELIXIR Interoperability Platform as a member of the IFB (Institut Français de Bioinformatique), the ELIXIR French Node (ELIXIR FR). She coordinates and reviews projects in the field of FAIR data, Data Management Plans and Recommended Interoperability Resources (RIR).

9.3 National initiatives

9.3.1 FEDER – SB-Server

Participants Marie-Dominique Devignes *(contact person)*, Bernard Maigret, Isaure Chauvot de Beauchêne, Sabeur Aridhi.

Project title: *Structural bioinformatics server*; PI: David Ritchie, Capsid (Inria Nancy – Grand Est); Value: 24 k€; Duration: 2015–2020. Description: This funding provides a small high performance computing server for structural bioinformatics research at the Inria Nancy – Grand Est centre.

9.3.2 ANR

FIGHT-HF

Participants Marie-Dominique Devignes *(contact person)*, Malika Smaïl-Tabbone *(contact person)*, Emmanuel Bresso, Bernard Maigret, Sabeur Aridhi, Kévin Dalleau, Claire Lacomblez, Gabin Personeni, Philippe Noel.

Project title: *Combattre l'insuffisance cardiaque : Projet de Recherche Hospitalo-Universitaire FIGHT-HF*; PI: Patrick Rossignol, Université de Lorraine (FHU-Cartage); Value: 9 M€ (Capsid and Orpailleur: 450 k€, approx); Duration: 2015–2020. Description: This "Investissements d'Avenir" project aims to discover novel mechanisms for heart failure and to propose decision support for precision medicine. The project has been granted 9 M€, and involves many participants from Nancy University Hospital Federation "CARTAGE". Marie-Dominique Devignes and Malika Smaïl-Tabbone are coordinating a work-package dedicated to network-based science, decision support and drug discovery for this project.

IFB

Participants Marie-Dominique Devignes *(contact person)*, Sabeur Aridhi, Isaure Chauvot de Beauchêne.

Project title: *Institut Français de Bioinformatique*; PI: Claudine Médigue and Jacques van Helden (CNRS UMS 3601); Value: 20 M€ (Capsid: 126 k€); Duration: 2014–2021. Description: The Capsid team is associated with the IFB (Institut Français de Bioinformatique), the French national network of bioinformatics platforms (http://www.france-bioinformatique.fr). The principal aim is to make bioinformatics skills and resources more accessible to French biology laboratories. Marie-Dominique Devignes is coordinating with Alban Gaignard the Interoperability task in the Integrative Bioinformatics Workpackage.

9.4 Regional initiatives

Amina Ahmed-Nacer benefited from a post-doctoral fellowship co-funded by the Region Grand-Est (Soutien Jeunes Chercheurs) and the Faculty Hospital (CHRU) of Nancy.

10 Dissemination

10.1 Promoting Scientific Activities

10.1.1 Scientific Events: Organisation

Member of the organizing committees Isaure Chauvot de Beauchêne is a member of the Organizing Committees of the annual winter school AlgoSB (Algorithms for Structural Biology).

Isaure Chauvot de Beauchêne was a member of the Organizing Committees of the Structural Bioinformatics symposium at JOBIM2020.

10.1.2 Scientific Events: Selection

Member of the conference program committees Isaure Chauvot de Beauchêne is a member of the Conference Program Committees of JOBIM2021 (Journéees Ouvertes en Biologie, Informatique et Mathématiques)

Marie-Dominique Devignes is a member of the Conference Program Committees of BIBM 2020, ACM-BCB 2020, IWBBIO 2020.

10.1.3 Journal

Member of the editorial boards Editorial board of Intelligent Data Analysis (Sabeur Aridhi)

Editor of the BTI (Bioinformatics and Translational Informatics) Section in the IMIA Yearbook 2020 (Malika Smaïl-Tabbone)

Reviewer - reviewing activities Marie-Dominique Devignes was reviewer for Nucleic Acids Research, Bionformatics, Journal of Biomedical Semantics, Frontiers and Biomarkers.

Isaure Chauvot de Beauchêne was reviewer for Computational and Structural Biotechnology Journal, Biomolecules,PROTEINS: Structure, Function, and Bioinformatics.

10.1.4 Scientific Expertise

Marie-Dominique Devignes reviewed a grant application for the Netherlands eScience Center (eTEC call 2020 : Innovative eScience Technologies)

10.1.5 Research Administration

Marie-Dominique Devignes is responsible for the transverse Axis Numeric Health at the LORIA.
Malika Smaïl-Tabbone and Isaure Chauvot de Beauchêne are members of the Conseil de Pôle AM2I at the Université de Lorraine.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Sabeur Aridhi and Malika Smaïl-Tabbone are enseignants-chercheurs with a full service. Sabeur Aridhi is responsible for the major in IAMD (Ingénierie et Applications des Masses de Données) at TELECOM Nancy (Université de Lorraine),
- Marie-Dominique Devignes teaches about 34h at Telecom Nancy (1A) and 10h in the Cursus Master Ingenieur at the Université de Lorraine.
- Isaure Chauvot de Beauchêne and Sjoerd de Vries organised and gave an online course on "Macromolecular Docking with ATTRACT", that took place in 4 sessions of 2 hours, between 8/03/2020 and 13/05/2020.

10.2.2 Supervision

- PhD in progress: Kévin Dalleau, *Complex graph analysis for classification: application to disease nosography,* 01/12/2016, Malika Smaïl-Tabbone, Miguel Couceiro.
- PhD in progress: Bishnu Sarker, *Developing distributed graph-based approaches for large-scale protein function annotation and knowledge discovery,* 01/11/2017, Marie-Dominique Devignes, Sabeur Aridhi.
- PhD in progress: Antoine Moniot, *Modeling protein / nucleic acid complexes by combinatorial structural fragment assembly,* 01/11/2018, Yann Guermeur, Isaure Chauvot de Beauchêne.
- PhD in progress: Athénaïs Vaginay, *Model selection and analysis for biological networks: use of domain knowledge and application to networks disturbed in diseases,* 01/11/2018, Taha Boukhobza, Malika Smaïl-Tabbone.
- PhD in progress: Anna Kravchenko, *Fragment-based modeling of protein-RNA complexes for protein design,* 01/10/2019, Malika Smaïl-Tabbone, Isaure Chauvot de Beauchêne.
- PhD in progress: Hrishikesh Dhondge, *A new knowledge base for modeling and design of RNA-binding proteins,* 01/10/2019, Marie-Dominique Devignes, Isaure Chauvot de Beauchêne.
- PhD in progress: Diego Amaya Ramirez, *HLA genetic system and organ transplantation: understanding the basics of immunogenicity to improve donor / receptor compatibility when assigning grafts to recipients,* 01/10/2019, Marie-Dominique Devignes, Jean-Luc Taupin.
- PhD in progress: Kamrul Islam, *Distributed link prediction in large complex graphs: application to biomolecule interactions,* 01/11/2019, Malika Smail-Tabbone, Sabeur Aridhi.

10.2.3 Juries

- Malika Smaïl-Tabbone was member of the PhD jury of Carlos Bermejo Das Neves (Université Strasbourg)

10.3 Popularization

10.3.1 Internal or External Inria Responsibilities

Malika Smaïl-Tabbone was a member of the national Inria SRP/ARP jury in 2020.

11 Scientific production

11.1 Major publications

[1] S. Z. Alborzi, M.-D. Devignes and D. W. Ritchie. 'ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains'. In: *BMC Bioinformatics* 18.1 (Dec. 2017), p. 107. DOI: 10.1186/s12859-017-1519-x. URL: https://hal.inria.fr/hal-01466842.

[2] S. Z. Alborzi, D. Ritchie and M.-D. Devignes. 'Computational Discovery of Direct Associations between GO terms and Protein Domains'. In: *BMC Bioinformatics* 19.Suppl 14 (Nov. 2018), p. 413. DOI: 10.1186/s12859-018-2380-2. URL: https://hal.inria.fr/hal-01777508.

[3] K. Dalleau, M. Couceiro and M. Smaïl-Tabbone. 'Unsupervised Extra Trees: a stochastic approach to compute similarities in heterogeneous data.' In: *International Journal of Data Science and Analytics* (Mar. 2020). DOI: 10.1007/s41060-020-00214-4. URL: https://hal.inria.fr/hal-01982232.

[4] A. W. Ghoorah, M.-D. Devignes, M. Smaïl-Tabbone and D. Ritchie. 'KBDOCK 2013: A spatial classification of 3D protein domain family interactions'. In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. 389–395. URL: https://hal.inria.fr/hal-00920612.

[5] A. W. Ghoorah, M.-D. Devignes, M. Smaïl-Tabbone and D. Ritchie. 'Protein Docking Using Case-Based Reasoning'. In: *Proteins* 81.12 (Oct. 2013), pp. 2150–2158. DOI: 10.1002/prot.24433. URL: https://hal.inria.fr/hal-00880341.

[6] A. W. Ghoorah, M.-D. Devignes, M. Smaïl-Tabbone and D. Ritchie. 'Spatial clustering of protein binding sites for template based protein docking'. In: *Bioinformatics* 27.20 (Aug. 2011), pp. 2820–2827. DOI: 10.1093/bioinformatics/btr493. URL: https://hal.inria.fr/inria-00617921.

[7] T. V. Hoang, X. Cavin and D. Ritchie. 'gEMfitter: A highly parallel FFT-based 3D density fitting tool with GPU texture memory acceleration'. In: *Journal of Structural Biology* (Sept. 2013). DOI: 10.1016/j.jsb.2013.09.010. URL: https://hal.inria.fr/hal-00866871.

[8] G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes and D. Ritchie. 'HexServer: an FFT-based protein docking server powered by graphics processors'. In: *Nucleic Acids Research* 38 (May 2010), W445–W449. DOI: 10.1093/nar/gkq311. URL: https://hal.inria.fr/inria-00522712.

[9] D. Ritchie. 'Calculating and scoring high quality multiple flexible protein structure alignments'. In: *Bioinformatics* 32.17 (May 2016), pp. 2650–2658. DOI: 10.1093/bioinformatics/btw300. URL: https://hal.inria.fr/hal-01371083.

[10] D. W. Ritchie and V. Venkatraman. 'Ultra-fast FFT protein docking on graphics processors'. In: *Bioinformatics* 26.19 (Aug. 2010), pp. 2398–2405. DOI: 10.1093/bioinformatics/btq444. URL: https://hal.inria.fr/inria-00537988.

[11] D. W. Ritchie and S. Grudinin. 'Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry'. In: *Journal of Applied Crystallography* 49.1 (Feb. 2016), pp. 158–167. DOI: 10.1107/S1600576715022931. URL: https://hal.inria.fr/hal-01261402.

[12] M. E. Ruiz Echartea, I. Chauvot de Beauchêne and D. Ritchie. 'EROS-DOCK: Protein-Protein Docking Using Exhaustive Branch-and-Bound Rotational Search'. In: *Bioinformatics* 35.23 (2019), pp. 5003–5010. DOI: 10.1093/bioinformatics/btz434. URL: https://hal.archives-ouvertes.fr/hal-02269812.

[13] M. E. Ruiz Echartea, D. Ritchie and I. Chauvot de Beauchêne. 'Using Restraints in EROS-Dock Improves Model Quality in Pairwise and Multicomponent Protein Docking'. In: *Proteins - Structure, Function and Bioinformatics* 88.8 (Aug. 2020), pp. 1121–1128. DOI: 10.1002/prot.25959. URL: https://hal.archives-ouvertes.fr/hal-02930827.

[14] B. Sarker, D. Ritchie and S. Aridhi. *GrAPFI: Graph Based Inference for Automatic Protein Function Annotation*. ECCB 2018 - 17th European Conference on Computational Biology. Poster. Sept. 2018. URL: https://hal.inria.fr/hal-01876907.

[15] B. Sarker, D. Ritchie and S. Aridhi. 'GrAPFI: predicting enzymatic function of proteins from domain similarity graphs'. In: *BMC Bioinformatics* (Apr. 2020). This work is dedicated to the memory of David W. Ritchie, who recently passed away. DOI: 10.1186/s12859-020-3460-7. URL: https://hal.inria.fr/hal-03022601.

[16] B. Sarker, D. W. Ritchie and S. Aridhi. 'Exploiting Complex Protein Domain Networks for Protein Function Annotation'. In: *Complex Networks 2018 - 7th International Conference on Complex Networks and Their Applications*. Cambridge, United Kingdom, Dec. 2018. URL: https://hal.inria.fr/hal-01920595.

11.2 Publications of the year

International journals

[17] S. Aridhi, J. Macedo, E. M. Nguifo and K. Zeitouni. 'Special issue on "Advances on Large Evolving Graphs"'. In: *Future Generation Computer Systems* 110 (Sept. 2020), p. 310. DOI: 10.1016/j.future.2020.04.023. URL: https://hal.inria.fr/hal-03025790.

[18] E. Bresso, A. Furlan, P. Noel, V. Leroux, F. Maina, R. Dono and B. Maigret. 'Large-Scale Virtual Screening Against the MET Kinase Domain Identifies a New Putative Inhibitor Type'. In: *Molecules* 25.4 (19th Feb. 2020), p. 938. DOI: 10.3390/molecules25040938. URL: https://hal.inria.fr/hal-03029061.

[19] P. Couvineau, H. de Almeida, V. Leroux, B. Roques, B. Maigret, C. Llorens-Cortes and X. Iturrioz. 'Structural insight into the catalytic mechanism and inhibitor binding of aminopeptidase A'. In: *Biochemical Journal* 477.21 (13th Nov. 2020), pp. 4133–4148. DOI: 10.1042/BCJ20200307. URL: https://hal.inria.fr/hal-03029058.

[20] K. Dalleau, M. Couceiro and M. Smaïl-Tabbone. 'Unsupervised Extra Trees: a stochastic approach to compute similarities in heterogeneous data.' In: *International Journal of Data Science and Analytics*. Advances in Intelligent Data Analysis {XVIII} - 18th International Symposium on Intelligent Data Analysis, IDA 2020 Lecture Notes in Computer Science.12080 (31st Mar. 2020), pp. 132–144. DOI: 10.1007/s41060-020-00214-4. URL: https://hal.inria.fr/hal-01982232.

[21] J. P. Ferreira, A. Pizard, J.-L. Machu, E. Bresso, H.-P. Brunner-LaRocca, N. Girerd, C. Leroy, A. González, J. Díez, S. Heymans, M.-D. Devignes, P. Rossignol and F. Zannad. 'Plasma protein biomarkers and their association with mutually exclusive cardiovascular phenotypes: the FIBRO-TARGETS case–control analyses'. In: *Clinical Research in Cardiology* 109.1 (2020), pp. 22–33. DOI: 10.1007/s00392-019-01480-4. URL: https://hal.inria.fr/hal-02138814.

[22] N. Girerd, E. Bresso, M.-D. Devignes and P. Rossignol. 'Insulin-like growth factor binding protein 2: A prognostic biomarker for heart failure hardly redundant with natriuretic peptides'. In: *International Journal of Cardiology* 300 (Feb. 2020), pp. 252–254. DOI: 10.1016/j.ijcard.2019.11.100. URL: https://hal.univ-lorraine.fr/hal-02517502.

[23] F. Inizan, M. Hanna, M. Stolyarchuk, I. Chauvot de Beauchêne and L. Tchertanov. 'The First 3D Model of the Full-Length KIT Cytoplasmic Domain Reveals a New Look for an Old Receptor'. In: *Scientific Reports* (25th Mar. 2020). DOI: 10.1038/s41598-020-62460-7. URL: https://hal.archives-ouvertes.fr/hal-03000379.

[24] J. Legrand, R. Gogdemir, C. Bousquet, K. Dalleau, M.-D. Devignes, W. Digan, C.-J. Lee, N.-C. Ndiaye, N. Petitpain, P. Ringot, M. Smaïl-Tabbone, Y. Toussaint and A. Coulet. 'PGxCorpus, a manually annotated corpus for pharmacogenomics'. In: *Scientific Data* 7.3 (2nd Jan. 2020). DOI: 10.1038/s41597-019-0342-9. URL: https://hal.inria.fr/hal-02547607.

[25] M. E. Ruiz Echartea, D. Ritchie and I. Chauvot de Beauchêne. 'Using Restraints in EROS-Dock Improves Model Quality in Pairwise and Multicomponent Protein Docking'. In: *Proteins - Structure, Function and Bioinformatics* 88.8 (Aug. 2020), pp. 1121–1128. DOI: 10.1002/prot.25959. URL: https://hal.archives-ouvertes.fr/hal-02930827.

[26] B. Sarker, D. Ritchie and S. Aridhi. 'GrAPFI: predicting enzymatic function of proteins from domain similarity graphs'. In: *BMC Bioinformatics* (29th Apr. 2020). DOI: 10.1186/s12859-020-3460-7. URL: https://hal.inria.fr/hal-03022601.

[27] M. Smaïl-Tabbone and B. Rance. 'Contributions from the 2019 Literature on Bioinformatics and Translational Informatics'. In: *IMIA Yearbook of Medical Informatics* 29.01 (21st Aug. 2020), pp. 188–192. DOI: 10.1055/s-0040-1702002. URL: https://hal.inria.fr/hal-03144080.

[28] S. Stienen, J. P. Ferreira, M. Kobayashi, G. Preud'homme, D. Dobre, J.-L. Machu, K. Duarte, E. Bresso, M.-D. Devignes, N. López Andrés, N. Girerd, S. Aakhus, G. Ambrosio, H.-P. Brunner-La Rocca, R. Fontes-Carvalho, A. G. Fraser, L. van Heerebeek, S. Heymans, G. de Keulenaer, P. Marino, K. McDonald, A. Mebazaa, Z. Papp, R. Raddino, C. Tschöpe, W. Paulus, F. Zannad and P. Rossignol. 'Enhanced clinical phenotyping by mechanistic bioprofiling in heart failure with preserved ejection fraction: insights from the MEDIA-DHF study (The Metabolic Road to Diastolic Heart Failure)'. In: *Biomarkers* 25.2 (17th Feb. 2020), pp. 201–211. DOI: 10.1080/1354750X.2020.1727015. URL: https://hal.univ-lorraine.fr/hal-02732968.

International peer-reviewed conferences

[29] A. Bazin, M. Couceiro, M.-D. Devignes and A. Napoli. 'Explaining Multicriteria Decision Making with Formal Concept Analysis'. In: Concept Lattices and Applications 2020. Vol. 2668. CEUR Workshop Proceedings. Tallinn, Estonia, 29th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02909383.

[30] K. Dalleau, M. Couceiro and M. Smaïl-Tabbone. 'Computing Vertex-Vertex Dissimilarities Using Random Trees: Application to Clustering in Graphs'. In: IDA 2020 -18th International Symposium on Intelligent Data Analysis. Vol. 12080. Lecture Notes in Computer Science. Konstanz / Virtual, Germany: https://ida2020.org/, 27th June 2020, pp. 132–144. DOI: 10.1007/978-3-030-44584-3_11. URL: https://hal.archives-ouvertes.fr/hal-02864678.

[31] A. Moniot, R. Roy, Y. Guermeur and I. Chauvot de Beauchêne. 'Docking of RNA Hairpin on Protein Using a Fragment-Based Method'. In: *Actes JOBIM 2020*. JOBIM. Montpellier, France, 1st July 2020. URL: https://hal.archives-ouvertes.fr/hal-02927185.

[32] P. Monnin, E. Bresso, M. Couceiro, M. Smaïl-Tabbone, A. Napoli and A. Coulet. 'Tackling scalability issues in mining path patterns from knowledge graphs: a preliminary study'. In: 1st international conference "Algebras, graphs and ordered sets" (Algos 2020). Proceedings of the 1st International Conference on Algebras, Graphs and Ordered Sets (ALGOS 2020). Nancy, France: https://algos2020.loria.fr/, 26th Aug. 2020. URL: https://hal.inria.fr/hal-02913224.

[33] B. Sarker, N. Khare, M.-D. Devignes and S. Aridhi. 'Graph Based Automatic Protein Function Annotation Improved by Semantic Similarity'. In: IWBBIO 2020 - 8th International Work-Conference on Bioinformatics and Biomedical Engineering. Vol. 12108. GRANADA, Spain, 30th Apr. 2020, pp. 261–272. DOI: 10.1007/978-3-030-45385-5_24. URL: https://hal.inria.fr/hal-03025827.

Scientific book chapters

[34] J. Lima, B. Maigret, D. Fernandez, J. Decloquement, D. Pinho, E. V. Albuquerque, M. Rodrigues and N. Martins. 'Searching in Silico Novel Targets for Specific Coffee Rust Disease Control'. In: *Lecture Notes in Computer Science, vol 11347. Springer*. 29th Apr. 2020, pp. 109–115. DOI: 10.1007/978-3-030-46417-2_10. URL: https://hal.inria.fr/hal-03029852.

Reports & preprints

[35] C. Bellanger, J. Hung, N. Juniarta, V. Leroux, B. Maigret and A. Napoli. *Chemoinformatics approaches to help antibacterial discovery*. Inria Nancy - Grand Est, 22nd May 2020. URL: https://hal.inria.fr/hal-02615395.

[36] W. Inoubli, S. Aridhi, H. ·. Mezni, M. Maddouri and E. Mephu Nguifo. *A Distributed and Incremental Algorithm for Large-Scale Graph Clustering*. 10th June 2020. URL: https://hal.inria.fr/hal-02190913.

Other scientific publications

[37] S. Gulberti, I. Bertin-Jung, I. Chauvot de Beauchêne, C. Valencia-Schmitt, P. Villa, B. Maigret and S. Fournel-Gigleux. *Vers un nouveau traitement des mucopolysaccharidoses ? Recherche d'inhibiteurs de la β1,4-galactosyltransférase 7 (β4GalT7) par des approches combinées de criblage expérimental et virtuel*. Strasbourg, France, 28th Feb. 2020. URL: https://hal.univ-lorraine.fr/hal-02946974.

11.3 Cited publications

[38] C. E. Alvarez-Martinez and P. J. Christie. 'Biological diversity of prokaryotic type IV secretion systems'. In: *Microbiology and Molecular Biology Reviews* 73 (2011), pp. 775–808.

[39] M. Baaden and S. R. Marrink. 'Coarse-grained modelling of protein-protein interactions'. In: *Current Opinion in Structural Biology* 23 (2013), pp. 878–886.

[40] A. Berchanski and M. Eisenstein. 'Construction of molecular assemblies via docking: modeling of tetramers with D_2 symmetry'. In: *Proteins* 53 (2003), pp. 817–829.

[41] P. Bork, L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee and E. M. Marcotte. 'Protein interaction networks from yeast to human'. In: *Current Opinion in Structural Biology* 14 (2004), pp. 292–299.

[42] I. J. Chauvot De Beauchene, S. J. De Vries and M. J. Zacharias. *Fragment-based modeling of protein-bound ssRNA*. ECCB 2016: The 15th European Conference on Computational Biology. Poster. Sept. 2016. URL: https://hal.archives-ouvertes.fr/hal-01573352.

[43] I. Chauvot de Beauchêne, S. J. De Vries and M. Zacharias. 'Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins'. In: *Nucleic Acids Research* (June 2016). DOI: 10.1093/nar/gkw328. URL: https://hal.archives-ouvertes.fr/hal-01505862.

[44] S. J. De Vries, I. Chauvot de Beauchêne, C. E. M. Schindler and M. Zacharias. 'Cryo-EM Data Are Superior to Contact and Interface Information in Integrative Modeling'. In: *Biophysical Journal* (Feb. 2016). DOI: 10.1016/j.bpj.2015.12.038. URL: https://hal.archives-ouvertes.fr/hal-01505863.

[45] M.-D. Devignes, S. Benabderrahmane, M. Smaïl-Tabbone, A. Napoli and O. Poch. 'Functional classification of genes using semantic distance and fuzzy clustering approach: Evaluation with reference sets and overlap analysis'. In: *international Journal of Computational Biology and Drug Design. Special Issue on: "Systems Biology Approaches in Biological and Biomedical Research"* 5.3/4 (2012), pp. 245–260. URL: https://hal.inria.fr/hal-00734329.

[46] S. E. Dobbins, V. I. Lesk and M. J. E. Sternberg. 'Insights into protein flexibility: The relationship between normal modes and conformational change upon protein–protein docking'. In: *Proceedings of National Academiy of Sciences* 105.30 (2008), pp. 10390–10395.

[47] M. El Houasli, B. Maigret, M.-D. Devignes, A. W. Ghoorah, S. Grudinin and D. Ritchie. 'Modeling and minimizing CAPRI round 30 symmetrical protein complexes from CASP-11 structural models'. In: *Proteins: Structure, Function, and Genetics*. Special Issue: Sixth Meeting on the Critical Assessment of Predicted Interactions 85.3 (Mar. 2017), pp. 463–469. DOI: 10.1002/prot.25182. URL: https://hal.inria.fr/hal-01388654.

[48] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus. 'Knowledge Discovery in Databases: An Overview'. In: *AI Magazine* 13 (1992), pp. 57–70.

[49] R. Fronzes, E. Schäfer, L. Wang, H. R. Saibil, E. V. Orlova and G. Waksman. 'Structure of a type IV secretion system core complex'. In: *Science* 323 (2011), pp. 266–268.

[50] R. A. Goldstein. 'The structure of protein evolution and the evolution of proteins structure'. In: *Current Opinion in Structural Biology* 18 (2008), pp. 170–177.

[51] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue and R. Apweiler. 'The HUPO PSI's Molecular Interaction format – a community standard for the representation of protein interaction data'. In: *Nature Biotechnology* 22.2 (2004), pp. 177–183.

[52] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole and S. R. Marrink. 'The power of coarse graining in biomolecular simulations'. In: *WIRES Comput. Mol. Sci.* 4 (2013), pp. 225–248. URL: http://dx.doi.org/10.1002/wcms.1169.

[53] J. D. Jackson. *Classical Electrodynamics*. New York: Wiley, 1975.

[54] P. J. Kundrotas, Z. W. Zhu and I. A. Vakser. 'GWIDD: Genome-wide protein docking database'. In: *Nucleic Acids Research* 38 (2010), pp. D513–D517.

[55] M. Lensink and S. J. Wodak. 'Docking and scoring protein interactions: CAPRI 2009'. In: *Proteins* 78 (2010), pp. 3073–3084.

[56] A. May and M. Zacharias. 'Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking'. In: *Proteins* 70 (2008), pp. 794–809.

[57] I. H. Moal and P. A. Bates. 'SwarmDock and the Use of Normal Modes in Protein-Protein Docking'. In: *International Journal of Molecular Sciences* 11.10 (2010), pp. 3623–3648.

[58] C. Morris. 'Towards a structural biology work bench'. In: *Acta Crystallographica* PD69 (2013), pp. 681–682.

[59] D. Mustard and D. Ritchie. 'Docking essential dynamics eigenstructures'. In: *Proteins: Structure, Function, and Genetics* 60 (2005), pp. 269–274. DOI: 10.1002/prot.20569. URL: https://hal.inria.fr/inria-00434271.

[60] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. L. Brinkman, G. Cesareni, A. Chatraryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. W. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stümpflen, M. Tyers, P. Uetz, I. Xenarios and H. Hermjakob. 'Protein interaction data curation: the International Molecular Exchange (IMEx) consortium'. In: *Nature Methods* 9.4 (2012), pp. 345–350.

[61] A. Özgur, Z. Xiang, D. R. Radev and Y. He. 'Mining of vaccine-associated IFN-γ gene interaction networks using the Vaccine Ontology'. In: *Journal of Biomedical Semantics* 2 (Suppl 2) (2011), S8.

[62] B. Pierce, W. Tong and Z. Weng. 'M-ZDOCK: A Grid-Based Approach for C_n Symmetric Multimer Docking'. In: *Bioinformatics* 21.8 (2005), pp. 1472–1478.

[63] D. Ritchie. 'Recent Progress and Future Directions in Protein-Protein Docking'. In: *Current Protein and Peptide Science* 9.1 (Feb. 2008), pp. 1–15. DOI: 10.2174/138920308783565741. URL: https://hal.inria.fr/inria-00434268.

[64] D. Ritchie and G. J. Kemp. 'Protein docking using spherical polar Fourier correlations'. In: *Proteins: Structure, Function, and Genetics* 39 (2000), pp. 178–194. URL: https://hal.inria.fr/inria-00434273.

[65] D. Ritchie, D. Kozakov and S. Vajda. 'Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational FFT generating functions'. In: *Bioinformatics* 24.17 (June 2008), pp. 1865–1873. DOI: 10.1093/bioinformatics/btn334. URL: https://hal.inria.fr/inria-00434264.

[66] A. Rivera-Calzada, R. Fronzes, C. G. Savva, V. Chandran, P. W. Lian, T. Laeremans, E. Pardon, J. Steyaert, H. Remaut, G. Waksman and E. V. Orlova. 'Structure of a bacterial type IV secretion core complex at subnanometre resolution'. In: *EMBO Journal* 32 (2013), pp. 1195–1204.

[67] M. G. Saunders and G. A. Voth. 'Coarse-graining of multiprotein assemblies'. In: *Current Opinion in Structural Biology* 22 (2012), pp. 144–150.

[68] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov and H. J. Wolfson. 'Geometry-based flexible and symmetric protein docking'. In: *Proteins* 60.2 (2005), pp. 224–231.

[69] M. L. Sierk and G. J. Kleywegt. 'Déjà vu all over again: Finding and analyzing protein structure similarities'. In: *Structure* 12 (2004), pp. 2103–2011.

[70] N. Soler, E. Robert, I. Chauvot de Beauchêne, P. Monteiro, V. Libante, B. Maigret, J. Staub, D. W. Ritchie, G. Guédon, S. Payot, M.-D. Devignes and N. N. Leblond-Bourget. 'Characterization of a relaxase belonging to the MOBT family, a widespread family in Firmicutes mediating the transfer of ICEs'. In: *Mobile DNA* 10.1 (Dec. 2019), pp. 1–16. DOI: 10.1186/s13100-019-0160-9. URL: https://hal.inria.fr/hal-02138843.

[71] A. Vaginay, M. Smail-Tabbone and T. Boukhobza. 'Towards an automatic conversion from SBML core to SBML qual'. In: *JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques*. Présentation Poster. Nantes, France, July 2019. URL: https://hal.archives-ouvertes.fr/hal-02407443.

[72] V. Venkatraman and D. Ritchie. 'Flexible protein docking refinement using pose-dependent normal mode analysis'. In: *Proteins* 80.9 (June 2012), pp. 2262–2274. DOI: 10.1002/prot.24115. URL: https://hal.inria.fr/hal-00756809.

[73] A. B. Ward, A. Sali and I. A. Wilson. 'Integrative Structural Biology'. In: *Biochemistry* 6122 (2013), pp. 913–915.

[74] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano and B. Honig. 'Structure-based prediction of protein-protein interactions on a genome-wide scale'. In: *Nature* 490 (2012), pp. 556–560.