IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2019

Project-Team CAPSID

Computational Algorithms for Protein Structures and Interactions

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

RESEARCH CENTER
Nancy - Grand Est

THEME
Computational Biology

Table of contents

Project-Team CAPSID

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

The CAPSID team has lost its leader Dave Ritchie who passed away on September 15, 2019. Marie-Dominique Devignes is the new leader of the CAPSID team.

Keywords:

Computer Science and Digital Science:

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, Visitors, External Collaborators

Research Scientists

Marie-Dominique Devignes [Team leader, CNRS, Researcher, HDR]
Isaure Chauvot de Beauchêne [CNRS, Researcher]
Bernard Maigret [CNRS, Emeritus]
David Ritchie [Inria, Senior Researcher, until Sep 2019, HDR]

Faculty Members

Sabeur Aridhi [Université de Lorraine, Associate Professor]
Malika Smaïl-Tabbone [Université de Lorraine, Associate Professor, HDR]

Post-Doctoral Fellows

Amina Ahmed Nacer [Université de Lorraine, Post-Doctoral Fellow, from Jul 2019]
Dominique Mias Lucquin [Université de Lorraine, Post-Doctoral Fellow, from May 2019]

PhD Students

Wissem Inoubli [Université de Lorraine, ATER, from Oct 2019]
Diego Amaya Ramirez [Inria, PhD Student, from Oct 2019]
Kévin Dalleau [CNRS, PhD Student]
Hrishikesh Dhondge [CNRS, PhD Student, from Oct 2019]
Kamrul Islam [Université de Lorraine, PhD Student, from Oct 2019]
Anna Kravchenko [CNRS, PhD Student, from Oct 2019]
Antoine Moniot [Université de Lorraine, PhD Student]
Gabin Personeni [CNRS, PhD Student, until Mar 2019]
Maria Elisa Ruiz Echartea [Inria, PhD Student]
Bishnu Sarker [Inria, PhD Student]
Athenaïs Vaginay [Université de Lorraine, PhD Student]

Technical staff

Emmanuel Bresso [CNRS, Engineer]
Claire Lacomblez [CNRS, Engineer, until Nov 2019]
Philippe Noel [CNRS, Engineer, from Sep 2019]

Interns and Apprentices

Patricia Alves Silva [CNRS, until Jul 2019]
Camille Depenveiller [Inria, from Feb 2019 until Jul 2019]
Honey Ashok Jain [Université de Lorraine, until Jun 2019]
Navya Khare [Inria, from May 2019 until Jul 2019]
Eloi Massoulié [Université de Lorraine, from Jun 2019 until Jul 2019]
Floriane Odje [Université de Lorraine, from Apr 2019 until Jun 2019]
Karina Mayumi Sakita [CNRS, from Oct 2019]

Administrative Assistants

Antoinette Courrier [CNRS, Administrative Assistant]
Isabelle Herlich [Inria, Administrative Assistant]

External Collaborators

Taha Boukhobza [Université de Lorraine]
Sjoerd Jacob de Vries [INSERM]
Vincent Leroux [Univ Denis Diderot, until Jun 2019]

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 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 protein 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 [69], [56].

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 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.

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 and protein-protein 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 protein molecules in order to help better understand how their 3D structures relate to their biological function. In summary, the Capsid team is organized 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, RNA, and DNA 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 protein-protein 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 [67],

[48]. 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 [52], [70]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [39].

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 [43], 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) [46].

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 [49] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [58]. 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 [71]. 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, [3], [5] will be updated and extended using newly published DDIs. Methods for inferring new DDIs from existing protein-protein interactions (PPIs) will be developped. 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 organized 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 [11], [12].

3.2. Integrative Multi-Component Assembly and Modeling

3.2.1. Context

At the molecular level, each PPI is embodied by a physical 3D protein-protein interface. Therefore, if the 3D structures of a pair of interacting proteins 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 flexibility accurately during docking is very computationally expensive. This is due to the very large number of internal degrees of freedom in each protein, 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 protein 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 [60], 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 [51]

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

This equation demonstrates 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 [61]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [10], [7]. 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 [62], [53], 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], [61]. In particular, many protein complexes involve symmetric arrangements of one or more sub-units, and the presence of symmetry may be exploited to reduce the search space considerably [38], [59], [66]. For example, using our operator notation (in which $\widehat{R}$ and $\widehat{T}$ 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

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

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 [8], and was subsequently applied to several symmetrical complexes from the "CAPRI" blind docking experiment [45]. 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 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 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 [57], [44], [54], [55]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [68], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [4].

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 [37]. 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 [65]. Furthermore, this kind of CG model effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [50]. We are currently developing a CG scoring function for fast protein-protein docking and multi-component assembly. This work is part of the PhD project of Maria-Elisa Ruiz-Echartea [19], [64]. Beyond this PhD project, the CG scoring function will be exploited in all our docing projects, especially for RNA-Protein docking (see below).

3.2.5. Assembling Multi-Component Complexes and Integrative Structure Modeling

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 [42], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space. This work has made some progress during the PhD project of Maria Elisa Ruiz Echartea but still requires further efforts.

3.2.6. Protein-Nucleic Acids 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 modelling 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 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 [41], [40].

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 knowledge-based energy function that will be learnt by machine-learning methods from public structural data on ssRNA-protein complexes.

In the future, we will improve the combinatorial algorithm used for reassembling the docked fragments using experimental constraints and machine-learning approaches.

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 personalized 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 personalized 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 analyzing 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 a UL-funded interdisciplinary project 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 phenoptyped 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[29]. Our approaches will be driven by knowledge and 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, David Ritchie, Philippe Noël, Antoine Moniot, 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 [36]. 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 [47], [63]. 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, INRA) 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

characterized a new class of relaxases [21]. 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 - Nucleic Acids Interactions

Participants: Isaure Chauvot de Beauchêne [contact person], David Ritchie, Dominique Mias-Lucquin, Antoine Moniot, Honey Jain, Anna Kravchenko, Hrishikesh Dhondge, Malika Smaïl-Tabbone, Marie-Dominique Devignes.

The second application domain of Axis 2 concerns protein-nucleic acids interactions. We need to assess and optimize our new algorithms on concrete protein-nucleic acids complexes in close collaboration with external partners coming from the experimental field of structural biology. To facilitate such collaborations, we will have to create 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 order to extend this approach to partially structured RNA molecules, we have built an automated pipeline to create (i) libraries of RNA fragments with arbitrary characteristics such as secondary structure, and (ii) testing benchmarks for applying these libraries to docking assays.

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

In the future, we will tackle a defined group of RNA-binding proteins containing RNA-Recognition Motif (RRM) domains. We will study existing and predicted complexes between various types of RRMs and various RNA sequences with computational methods in order to calculate CG force-field energy and to help design new synthetic proteins with targeted RNA specificity. This is the goal of the ITN RNAct project and it will require the construction of a dedicated database equipped with querying and analysis facilities, including machine-learning approaches, as well as many interactions within the ITN RNAct consortium.

5. Highlights of the Year

5.1. Highlights of the Year

Malika Smaïl-Tabbone was invited with Bastien Rance to coordinate the selection of the best contributions from 2018 literature on Bioinformatics and Translational Informatics for the 2019 IMIA YearBook of Medical Informatics [20].

Bishnu Sarker (PhD student) obtained a DrEAM fellowship from Lorraine Université d'Excellence for a 3-month internship at the MILA (Machine Learning Laboratory of the University of Montreal and University of Quebec) in Montreal.

6. New Software and Platforms

6.1. lib3Dmol

Library in Rust for manipulating 3D representations of molecules

KEYWORDS: 3D modeling - Proteins - Molecules - Rust

FUNCTIONAL DESCRIPTION: The lib3Dmol library can be called by programs written in Rust for 3D modelling of biomolecules and their interactions.

RELEASE FUNCTIONAL DESCRIPTION: The 0.2.0 version can be used with any type of biomolécule.

- Contact: Philippe Noel
- URL: http://mbi.loria.fr

6.2. QRMSDmap

Parallelized computation of RMSD map of molecular structures after 3D alignment based on the quaternion method.

KEYWORDS: Molecules - RMSD - Rust - Bioinformatics

FUNCTIONAL DESCRIPTION: This program allows fast computing of 3D alignments and 3D distances on a large number of biomolecular structures.

RELEASE FUNCTIONAL DESCRIPTION: This 2.3.2 version improves CPU parallelization and decreases memory consumption.

- Contact: Philippe Noel
- URL: http://mbi.loria.fr

6.3. EROS-DOCK

Exhaustive Rotational Search using Branch-and-Bound algorithm for rigid docking

KEYWORDS: 3D modeling - Proteins - Docking

FUNCTIONAL DESCRIPTION: EROS-DOCK is a protein-protein docking program for Linux. It takes in input the 3D structures of two proteins in PDB format, and gives as output a list of transformation matrices describing the most probable relative positions of the two proteins in nature, together with a score (approximation of their binding energy for that position). On a modern workstation, docking times is in the order of few hours for a blind global search. The user can also provide nowledge of particular contact points at the surface of each protein, which accelerates the pruning of the solutions space. The underlying algorithm uses a pi-ball representation of the rotational 3D space, to accelerate the search for close-fitting orientations of the two molecules by a branch-and-bound technique.

- Contact: Isaure Chauvot de Beauchêne
- URL: https://erosdock.loria.fr

6.4. NAFRAGDB

Databases of nucleic acids fragments bound to proteins

KEYWORDS: Structural Biology - Nucleic Acids - Data base

FUNCTIONAL DESCRIPTION: NAfragDB 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 customized queries, and (iv) the computation of statistics on sets of RNA/DNA - protein structures.

- Contact: Isaure Chauvot de Beauchêne

6.5. RNA-PDBComplete

Completing RNA structures in PDB files

KEYWORDS: Nucleic Acids - Structural Biology

FUNCTIONAL DESCRIPTION: PDBcomplete is a software and a webserver for the completion of missing atoms in an RNA structure provided in PDB format. PDBcomplete is capable of taking into account the presence of other molecules in the overall PDB structure to avoid atoms collisions. It uses as template an in-house library of mono-nucleotide libraries created with the NAfragDB tool.

- Contact: Isaure Chauvot de Beauchêne
- URL: https://pdbcomplete.loria.fr/

6.6. MBI platform for structural bioinformatics

Initiated during the previous CPER projects Intelligence Logicielle (1999-2005) and MISN: Modelisation, Interactions et Systèmes Numériques (2006-2013), the MBI platform (MBI = Modelling Biomolecules and their Interactions) is today part of the SMEC platform coordinated by MD Devignes and M Smaïl-Tabbone (SMEC: Simulation, Modélisation et Extraction de Connaissances), in the frame of the ongoing CPER projet ITM2P (Innovations Technologiques et Modélisation pour la Médecine Personnalisée ; 2015-2020). The MBI platform is composed of several HPC and storage servers that are shared between users mostly for structural bioinformatics usages. The MBI platform is part of the bioinformatic platform network of the French Institute of Bioinformatics (IFB ; http://www.france-bioinformatique.fr.

- Participants: Marie-Dominique Devignes [contact person], Isaure Chauvot de Beauchêne, Sjoerd de Vries, Antoine Moniot, Emmanuel Bresso, Philippe Noël, Patrice Ringot.
- URL: https://mbi.loria.fr

7. New Results

7.1. Axis 1 : New Approaches for Knowledge Discovery in Structural Databases

7.1.1. Biomedical Knowledge Discovery

Our collaboration with clinicians at the CHRU Nancy in the framework of the RHU FIGHT-HF program and of the Contrat d'Interface has lead to two publications demonstrating the added value of database and knowledge graph exploitation when analyzing observational or prospective cohorts. In a retrospective observational study, we have identified and characterized patient subgroups presenting stable or unstable positivity to anti-phospholipid antobodies assays [15]. In the European FibroTarget cohort study, we have contributed to the characterization of at-risk phenotypic groups using proteomic biomarkers [16].

Another application is carried out in collaboration with the Orpailleur Team and concerns the PraktikPharma ANR project. We aim at building explanations for severe drug side effects (such as drug-induced liver injury or severe cutaneous adverse reaction) from pharmacogenomics RDF graph (PGXlod). We obtained a podium abstract at the MedInfo 2019 conference for providing molecular characterization for unexplained adverse drug reactions using pharmacogenomics RDF graph (PGXlod) [30].

7.1.2. Stochastic Decision Trees for Similarity Computation

In the frame of Kévin Dalleau's PhD thesis, we have designed a method to compute similarities on unlabeled data using stochastic decision trees [32], [27]. The main idea of Unsupervised Extremely Randomized Trees (UET) is to randomly and iteratively split the data until a stopping criterion is met. Pairwise similarity values are computed based on the co-occurrence of samples in the leaves of each generated tree. We evaluate our method on synthetic and real-world datasets by comparing the mean similarities between samples with the same label and the mean similarities between samples with distinct labels. Empirical studies show that the method effectively gives distinct similarity values between samples belonging to distinct clusters, and gives indiscernible values when there is no cluster structure. We also assessed some interesting properties such as invariance under monotone transformations of variables and robustness to correlated variables and noise. Our

experiments show that the algorithm outperforms existing methods in some cases, and can reduce the amount of preprocessing needed with many real-world datasets. We extended the approach to the computation of pairwise similarity for graph nodes. The experimental results are competitive with state of the art methods. We are currently working on merging the two similarity methods (on attribute-value objects and on graph nodes) to attributed graphs where the nodes are described by attributes.

We plan to study the application of this pairwise similarity computation to quantify protein structural similarities. Two interesting problems will concern the representation of the protein structure and how to tackle extra constraints such as invariance under rotational and translational transformations.

7.1.3. Protein Annotation and Machine Learning

We have been involved in the 3rd international CAFA Challenge ("Critical Assessment of Functional Annotation") through our work on (i) domain functional annotation (Zia Alborzi's PhD thesis) and (ii) label propagation in graphs (Bishnu Sarker's PhD thesis). We were therefore contributors of the general report published this year [23].
As part of his PhD work, Bishnu Sarker developed and tested on UniProt/SwissProt a new method for functional annotation of proteins using domain embedding-based sequence classification [25].
Multiple Instance Learning (MIL) is a machine learning strategy that can be applied to sets of sequences describing organisms displaying a given property. The purpose here is to be able to classify a new organism with respect to this property based on its sequences and their similarity to the sequences of classified organisms. New MIL algorithms have been described and tested in the framework of a collaboration [26], [24]. Another collaborative work has lead to the development of a distributed algorithm for large-scale graph clustering [34].

7.2. Axis 2 : Integrative Multi-Component Assembly and Modeling

7.2.1. EROS-DOCK algorithm and its extensions

We have adapted our EROS-DOCK protein-protein docking software [35], [19] to account for experimental knowledge on the protein-protein interface to be modeled. Indeed, structural biology experiments can identify pairs of amino-acids from each protein in a protein-protein interface that are likely to be in close contact. This additional restraint is used to pre-prune the 3D rotational space of one protein toward another, by eliminating cones of rotations that cannot fulfill the distance between the two points at the protein surfaces. Using a single restraint permits to decrease the average execution time by at least 90 percent.

We also developed a new version of EROS-DOCK for multi-body docking (modeling assemblies of more than 2 proteins), using a combinatorial approach. We assembled trimers by docking in a first stage all possible combinations of pairs of proteins involved in the multi-body complex. Possible trimer solutions are assembled by fixing one protein, the "root-protein" (protein A, say) at the origin and by placing the other two around it using the transformations,T[AB] and T[AC], from the corresponding pairwise solution lists returned by EROS-DOCK. If the three transformations together form a near-native (biologically relevant) trimer structure, then it is natural to suppose that T[BC] should be found in the list of B-C pairwise solutions.

Both extensions of the EROS-DOCK algorithm reported last year and published early this year [19] have been presented by Maria-Elisa Ruiz Echartea at the 2019 CAPRI meeting in april 2019 (http://www.capri-docking.org/events/) and at the MASIM meeting in november 2019 [28]. These results are part of her PhD Thesis that was defended on december 18, 2019 (the thesis will soon be available on HAL). A paper describing EROS-DOCK adaptation to multi-body docking in under revision in *Proteins*.

7.2.2. Protein docking

The regular participation of the Capsid team to the CAPRI challenge is acknowledged through its contribution to the review published this year on CAPRI round 46 [17].

We also contributed to an evaluation of docking software performance in protein-glycosaminoglycan systems [22].

7.2.3. 3D modeling and virtual screening

We have built a 3D model by homology of a new class of relaxase involved in the horizontal transfer of DNA in a group of bacteria called Firmicutes [21].

We also built a 3D model of a chemosensory GPCR as a potential target to control a parasite in plants [13].

Virtual screening was applied on various targets in a re-purposing strategy and led to the discovery of small molecules active against invasive fungal disease [14], [18].

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. CPER – IT2MP

Participants: Marie-Dominique Devignes [contact person], Malika Smaïl-Tabbone, David Ritchie.

Project title: *Innovations Technologiques, Modélisation et Médecine Personnalisée*; PI: Faiez Zannad, Université de Lorraine (Inserm-CHU-UL). Value: 14.4 M€ ("SMEC" platform – Simulation, Modélisation, Extraction de Connaissances – coordinated by Capsid and Orpailleur teams for Inria Nancy – Grand Est, with IECL and CHRU Nancy: 860 k€, approx); Duration: 2015–2020. Description: The IT2MP project encompasses four interdisciplinary platforms that support several scientific pôles of the university whose research involves human health. The SMEC platform supports research projects ranging from molecular modeling and dynamical simulation to biological data mining and patient cohort studies.

8.1.2. LUE-FEDER – CITRAM

Participants: Marie-Dominique Devignes [contact person], Isaure Chauvot de Beauchêne, Bernard Maigret, Philippe Noël, Dominique Mias-Lucquin, Antoine Moniot, David Ritchie.

Project title: *Conception d'Inhibiteurs du Transfert de Résistances aux agents Anti-Microbiens: bio-ingénierie assistée par des approches virtuelles et numériques, et appliquée à une relaxase d'élément conjugatif intégratif*; PI: N. Leblond, Université de Lorraine (DynAMic, UMR 1128); Other partners: Chris Chipot, CNRS (LPCT, UMR 7565); Value: 200 k€ (Capsid: 80 k€); Duration: 2017–2018. Description: This project follows on from the 2016 PEPS project "MODEL-ICE". The aim is to investigate protein-protein interactions required for initiating the transfer of an ICE (Integrated Conjugative Element) from one bacterial cell to another one, and to develop small-molecule inhibitors of these interactions.

8.1.3. IMPACT GeenAge

Participant: Marie-Dominique Devignes [contact person].

The IMPACT project GeenAge (Lorraine Université d'Excellence) is composed of four axes dedicated to research in high-throughput molecular biology. The Capsid team is involved in a transversal axis for numerical sciences. In the frame of this project, Marie-Dominique Devignes co-supervises with Amedeo Napoli a post-doc hired by the Orpailleur team. She is also responsible with Thierry Bastogne (CRAN) and Anne Gegout-Petit (IECL) for creating a Center of Competencies in Artificial Intelligence and Health.

8.2. National Initiatives

8.2.1. FEDER – SB-Server

Participants: Marie-Dominique Devignes [contact person], Bernard Maigret, Isaure Chauvot de Beauchêne, Sabeur Aridhi, David Ritchie.

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.

8.2.2. ANR

8.2.2.1. 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 Noël, David Ritchie.

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 € 9M, and involves many participants from Nancy University Hospital's 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.

8.2.2.2. IFB

Participants: Marie-Dominique Devignes [contact person], Sabeur Aridhi, Isaure Chauvot de Beauchêne, David Ritchie.

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 a research node of 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.

8.3. European Initiatives

8.3.1. FP7 & H2020 Projects

8.3.1.1. H2020 ITN RNAct

Participants: Isaure Chauvot de Beauchêne [contact person], Marie-Dominique Devignes, Malika Smaïl-Tabbone, Hrishikesh Dhondge, Anna Kravchenko, David Ritchie.

Program: H2020 Innovative Training Network

Project acronym:RNAct

Project title: Enabling proteins with RNA recognition motifs for synthetic biology and bio-analytics

Duration: octobre 2018 - octobre 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

8.3.2. Informal European Partners

EBI: European Bioinformatics Institute, Maria Martin team (UK). We are working with the EBI team to validate and improve our graph-based approaches for protein function annotation.

ELIXIR: 3D-bioinfo Community. We participated in the creation of the new ELIXIR 3D-bioinfo 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 supports its formally recognised Communities by providing funding for workshops and short collaborative projects associated with the Community. More specifically, Isaure Chauvot de Beauchene is member of the sub-section "Tools to describe, analyze, annotate, and predict nucleic acid structures" of this community.

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

8.4. International Initiatives

8.4.1. TempoGraphs

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

Participants: Sabeur Aridhi (PI), Marie-Dominique Devignes, Malika Smaïl-Tabbone, Bishnu Sarker, Wissem Inoubli, Dave Ritchie.

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.

8.4.2. Inria Associate Teams Not Involved in an Inria International Labs

Project: FlexMol. Algorithms for Multiscale Macromolecular Flexibility:

Participants: Maria-Elisa Ruiz-Echartea, Dave Ritchie, Isaure Chauvot de Beauchêne.

Partners: Nano-D, ChaconLab team, Rocasolano Institute of Physical Chemistry (IQFR-CSIC), Madrid, Spain, as non-beneficiary associated lab.

Description: Developing representations of molecular flexibility at different scales, for the 3D modeling of multi-molecular assemblies.

8.4.3. Informal International Partners

Project: Characterization, expression and molecular modeling ofTRR1 and ALS3 proteins of Candida spp., as a strategy to obtain new drugs with action on yeasts involved in nosocomial infections. Participant: Bernard Maigret. Partner: State University of Maringá, Brasil. Publication: [14], [18].

Project: Fusarium graminearum target selection. Participant: Bernard Maigret. Partner: Embrapa Recursos Geneticos e Biotecnologia, Brasil. Publication: [13].

Project:The thermal shock HSP90 protein as a target for new drugs against paracoccidioidomicose. Participant: Bernard Maigret. Partner: Brasília University, Brasil.

Project:Protein-protein interactions for the development of new drugs. Participant: Bernard Maigret. Partner: Federal University of Goias, Brasil.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events: Organisation

- Sabeur Aridhi co-chaired the third international workshop on Advances in managing and mining large evolving graphs (LEG - https://leg-ecmlpkdd19.loria.fr/) held in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2019).
- Isaure Chauvot de Beauchêne and Marie-Dominique Devignes organised the first international Workshop (5 days) of the H2020-ITN project RNAct, "RRMs, RNA and RNAct" (http://rnact.eu/Workshop1/)
- Isaure Chauvot de Beauchêne organised the 3rd meeting (regional, 1 day, 25 pers.) of the Glyco-EST group. GlycoEst is an informal working group which was recently created to develop an interdisciplinary regional network of glyco-scientists.
- Isaure Chauvot de Beauchêne organised a lecture and practical course at the AlgoSB WinterSchool 14-21 january 2019 on *Predicting RNA-Protein Interactions*.

9.1.2. Scientific Events: Selection

- Members of the following Conference Program Committees : Joint ICML 2019 Workshop on Computational Biology, IWBBIO 2019, ACM-BCB 2019, BIBM 2019, SWAT4HCLS 2019, EGC 2019.

9.1.3. Journal

- Editorial board of Intelligent Data Analysis (Sabeur Aridhi), Scientific reports (David Ritchie).
- Reviewer for Nucl. Acids Research (Marie-Dominique Devignes).
- Contribution to the IMIA Yearbook of Medical Informatics, 2019 (Malika Smaïl-Tabbone, [20]

9.1.4. Leadership within the Scientific Community

- Isaure Chauvot de Beauchêne is co-founder of the 3D Bioinfo ELIXIR Community.
- Marie-Dominique Devignes and Malika Smaïl-Tabbone have been invited to participate in the INI-CRCT network which is the subnetwork of the F-CRIN project (French Clinical Research Investigation Network) dedicated to cardio-renal diseases. Their contribution is related to their expertise in machine learning and network science.

9.1.5. Scientific Expertise

- Marie-Dominique Devignes reviewed a grant application for FWO (Flanders Research Organization).
- Malika Smaïl-Tabbone and Marie-Dominique Devignes both reviewed grant applications for the ANR.

9.1.6. Research Administration

- Sabeur Aridhi is a member of the Inria Nancy Grand-Est CDT: Commission du Développement Technologique.

- Marie-Dominique Devignes was a member of the ComiPers at Inria Nancy Grand-Est: Commission for the evaluation of CORDI post-doc and CORDI-S doctoral application.
- Malika Smaïl-Tabbone is a member of the IES commission for the elaboration of the policy concerning Scientific Information and Edition Scientifique at Inria Nancy Grand-Est and at the LORIA.
- Dave Ritchie was a member of the CMI at the LORIA: Commission de Mention Informatique of the Université de Lorraine's IAEM doctoral school.

9.2. Teaching - Supervision - Juries

9.2.1. 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 teaches about 10h in the Cursus Master Ingenieur at the Université de Lorraine.

9.2.2. Supervision

- PhD: Maria Elisa Ruiz Echartea, *Multi-component protein assembly using distance constraints*. Université de Lorraine. Defense date : 18/12/2019 (Manuscript under revision, soon in HAL). David Ritchie, Isaure Chauvot de Beauchêne.
- 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, David Ritchie, Sabeur Aridhi.
- PhD in progress: Antoine Moniot, *Modeling protein / nucleic acid complexes by combinatorial structural fragment assembly,* 01/11/2018, David Ritchie, 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.

9.2.3. Juries

- Sabeur Aridhi was a member (examinator) of the PhD committee of Manel Zoghlami, Universiy of Clermont Auvergne, *Multiple instance learning approaches for ionizing-radiation-resistance prediction,* 20/12/2019.
- Sabeur Aridhi was a member (reviewer) of the PhD committee of Zekarias Tilahun Kefato, Universiy of Trento, *Network and Cascade Representation Learning Algorithms based on Information Diffusion Events,* 29/04/2019.

- Sabeur Aridhi was a member (reviewer) of the PhD committee of Nasrullah Sheikh, Universiy of Trento, *Network Representation Learning with Attributes and Heterogeneity,* 16/07/2019.
- Marie-Dominique Devignes was a member (reviewer) of the PhD committee of Manel Zoghlami, Universiy of Clermont Auvergne, *Multiple instance learning approaches for ionizing-radiation-resistance prediction,* 20/12/2019.

9.3. Popularization

9.3.1. Interventions

- Dominique Mias-Lucquin was co-organizer of the "Pint of Science" event, 21-22 may, 2019 (24 countries involved ; http://pintofscience.com).

10. Bibliography

Major publications by the team in recent years

[1] S. Z. ALBORZI, M.-D. DEVIGNES, D. W. RITCHIE. *ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains*, in "BMC Bioinformatics", December 2017, vol. 18, n° 1, 107 p. [*DOI :* 10.1186/S12859-017-1519-X], https://hal.inria.fr/hal-01466842

[2] S. Z. ALBORZI, D. RITCHIE, M.-D. DEVIGNES. *Computational Discovery of Direct Associations between GO terms and Protein Domains*, in "BMC Bioinformatics", November 2018, vol. 19, n° Suppl 14, 413 p. [*DOI :* 10.1186/S12859-018-2380-2], https://hal.inria.fr/hal-01777508

[3] A. W. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Spatial clustering of protein binding sites for template based protein docking*, in "Bioinformatics", August 2011, vol. 27, n° 20, pp. 2820-2827 [*DOI :* 10.1093/BIOINFORMATICS/BTR493], https://hal.inria.fr/inria-00617921

[4] A. W. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Protein Docking Using Case-Based Reasoning*, in "Proteins", October 2013, vol. 81, n° 12, pp. 2150-2158 [*DOI :* 10.1002/PROT.24433], https://hal.inria.fr/hal-00880341

[5] A. W. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *KBDOCK 2013: A spatial classification of 3D protein domain family interactions*, in "Nucleic Acids Research", January 2014, vol. 42, n° D1, pp. 389-395, https://hal.inria.fr/hal-00920612

[6] T. V. HOANG, X. CAVIN, D. RITCHIE. *gEMfitter: A highly parallel FFT-based 3D density fitting tool with GPU texture memory acceleration*, in "Journal of Structural Biology", September 2013 [*DOI :* 10.1016/J.JSB.2013.09.010], https://hal.inria.fr/hal-00866871

[7] G. MACINDOE, L. MAVRIDIS, V. VENKATRAMAN, M.-D. DEVIGNES, D. RITCHIE. *HexServer: an FFT-based protein docking server powered by graphics processors*, in "Nucleic Acids Research", May 2010, vol. 38, pp. W445-W449 [*DOI :* 10.1093/NAR/GKQ311], https://hal.inria.fr/inria-00522712

[8] D. W. RITCHIE, S. GRUDININ. *Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry*, in "Journal of Applied Crystallography", February 2016, vol. 49, n° 1, pp. 158-167 [*DOI :* 10.1107/S1600576715022931], https://hal.inria.fr/hal-01261402

[9] D. RITCHIE. *Calculating and scoring high quality multiple flexible protein structure alignments*, in "Bioinformatics", May 2016, vol. 32, n^{o} 17, pp. 2650-2658 [*DOI :* 10.1093/BIOINFORMATICS/BTW300], https://hal.inria.fr/hal-01371083

[10] D. W. RITCHIE, V. VENKATRAMAN. *Ultra-fast FFT protein docking on graphics processors*, in "Bioinformatics", August 2010, vol. 26, n^{o} 19, pp. 2398-2405 [*DOI :* 10.1093/BIOINFORMATICS/BTQ444], https://hal.inria.fr/inria-00537988

[11] B. SARKER, D. RITCHIE, S. ARIDHI. *GrAPFI: Graph Based Inference for Automatic Protein Function Annotation*, September 2018, ECCB 2018 - 17th European Conference on Computational Biology, Poster, https://hal.inria.fr/hal-01876907

[12] B. SARKER, D. W. RITCHIE, 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, December 2018, https://hal.inria.fr/hal-01920595

Publications of the year

Articles in International Peer-Reviewed Journals

[13] E. BRESSO, D. FERNANDEZ, D. X. AMORA, P. NOEL, A.-S. PETITOT, M.-E. LISEI DE SA, E. V. S. ALBUQUERQUE, E. DANCHIN, B. MAIGRET, N. F. MARTINS. *A Chemosensory GPCR as a Potential Target to Control the Root-Knot Nematode Meloidogyne incognita Parasitism in Plants*, in "Molecules", 2019, vol. 24, n^{o} 20, 3798 p. [*DOI :* 10.3390/MOLECULES24203798], https://hal.archives-ouvertes.fr/hal-02324816

[14] I. R. G. CAPOCI, D. R. FARIA, K. M. SAKITA, F. A. V. RODRIGUES-VENDRAMINI, P. D. S. BONFIM-MENDONÇA, T. C. A. BECKER, E. S. KIOSHIMA, T. I. E. SVIDZINSKI, B. MAIGRET. *Repurposing approach identifies new treatment options for invasive fungal disease*, in "Bioorganic Chemistry", March 2019, vol. 84, pp. 87-97 [*DOI :* 10.1016/J.BIOORG.2018.11.019], https://hal.inria.fr/hal-02151642

[15] J. DEVIGNES, M. SMAÏL-TABBONE, A. HERVÉ, G. CAGNINACCI, M.-D. DEVIGNES, T. LECOMPTE, S. ZUILY, D. WAHL. *Extended persistence of antiphospholipid antibodies beyond the twelve-week time interval: Association with baseline antiphospholipid antibodies titres*, in "International Journal of Laboratory Hematology", September 2019, vol. 41, n^{o} 6, pp. 726-730 [*DOI :* 10.1111/IJLH.13094], https://hal.inria.fr/hal-02395258

[16] J. P. FERREIRA, A. PIZARD, J.-L. MACHU, E. BRESSO, H.-P. BRUNNER-LAROCCA, N. GIRERD, L. CÉLINE, A. GONZÁLEZ, J. DÍEZ, S. HEYMANS, M.-D. DEVIGNES. *Plasma protein biomarkers and their association with mutually exclusive cardiovascular phenotypes: the FIBROTARGETS case–control analyses*, in "Clinical Research in Cardiology", April 2019, vol. 1 [*DOI :* 10.1007/S00392-019-01480-4], https://hal.inria.fr/hal-02138814

[17] M. LENSINK, G. BRYSBAERT, N. NADZIRIN, S. VELANKAR, R. A. CHALEIL, T. GERGURI, P. BATES, E. LAINE, A. CARBONE, S. GRUDININ, R. KONG, R. LIU, X. XU, H. SHI, S. CHANG, M. EISENSTEIN, A. KARCZYNSKA, C. CZAPLEWSKI, E. LUBECKA, A. LIPSKA, P. KRUPA, M. MOZOLEWSKA, Ł. GOLON, S. SAMSONOV, A. LIWO, S. CRIVELLI, G. PAGÈS, M. KARASIKOV, M. KADUKOVA, Y. YAN, S. HUANG, M. ROSELL, L. A. RODRÍGUEZ-LUMBRERAS, M. ROMERO-DURANA, L. DÍAZ-BUENO, J. FERNANDEZ-RECIO, C. CHRISTOFFER, G. TERASHI, W. SHIN, T. ADERINWALE, S. RAGHAVENDRA MADDHURI VENKATA SUBRAM, D. KIHARA, D. KOZAKOV, S. VAJDA, K. PORTER, D. PADHORNY, I. DESTA, D. BEGLOV, M. IGNATOV, S. KOTELNIKOV, I. MOAL, D. RITCHIE, I. CHAUVOT DE BEAUCHÊNE, B.

MAIGRET, M. E. R. ECHARTEA, D. BARRADAS-BAUTISTA, Z. CAO, L. CAVALLO, R. OLIVA, Y. CAO, Y. SHEN, M. BAEK, T. PARK, H. WOO, C. SEOK, M. BRAITBARD, L. BITTON, D. SCHEIDMAN-DUHOVNY, J. DAPKŪNAS, K. OLECHNOVIČ, Č. VENCLOVAS, P. J. KUNDROTAS, S. BELKIN, D. CHAKRAVARTY, V. BADAL, I. A. VAKSER, T. VREVEN, S. VANGAVETI, T. M. BORRMAN, Z. WENG, J. D. GUEST, R. GOWTHAMAN, B. G. PIERCE, X. XU, R. DUAN, L. QIU, J. HOU, B. RYAN MERIDETH, Z. MA, J. CHENG, X. ZOU, P. KOUKOS, J. ROEL-TOURIS, F. AMBROSETTI, C. GENG, J. SCHAARSCHMIDT, M. TRELLET, A. S. MELQUIOND, L. XUE, B. JIMÉNEZ-GARCÍA, C. NOORT, R. HONORATO, A. M. BONVIN, S. J. WODAK. *Blind prediction of homo- and hetero- protein complexes: The CASP13-CAPRI experiment*, in "Proteins - Structure, Function and Bioinformatics", October 2019, vol. 87, n^{o} 12, pp. 1200-1221 [*DOI :* 10.1002/PROT.25838], https://hal.inria.fr/hal-02320974

[18] F. A. V. RODRIGUES-VENDRAMINI, D. R. FARIA, G. S. ARITA, I. R. G. CAPOCI, K. M. SAKITA, S. M. CAPARROZ-ASSEF, T. C. A. BECKER, P. DE SOUZA BONFIM-MENDONÇA, M. S. FELIPE, T. I. E. SVIDZINSKI, B. MAIGRET, E. S. KIOSHIMA, P. SMALL. *Antifungal activity of two oxadiazole compounds for the paracoccidioidomycosis treatment*, in "PLoS Neglected Tropical Diseases", June 2019, vol. 13, n^{o} 6, e0007441 p. [*DOI :* 10.1371/JOURNAL.PNTD.0007441], https://hal.inria.fr/hal-02151638

[19] M. E. RUIZ ECHARTEA, I. CHAUVOT DE BEAUCHÊNE, D. RITCHIE. *EROS-DOCK: Protein-Protein Docking Using Exhaustive Branch-and-Bound Rotational Search*, in "Bioinformatics", 2019 [*DOI :* 10.1093/BIOINFORMATICS/BTZ434], https://hal.archives-ouvertes.fr/hal-02269812

[20] M. SMAÏL-TABBONE, M. SMAIL-TABBONE, B. RANCE. *Contributions from the 2018 Literature on Bioinformatics and Translational Informatics*, in "IMIA Yearbook of Medical Informatics", August 2019, vol. 28, n^{o} 01, pp. 190-193 [*DOI :* 10.1055/S-0039-1677945], https://hal.inria.fr/hal-02413623

[21] 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, 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", December 2019, vol. 10, n^{o} 1 [*DOI :* 10.1186/S13100-019-0160-9], https://hal.inria.fr/hal-02138843

[22] U. UCIECHOWSKA-KACZMARZYK, I. CHAUVOT DE BEAUCHÊNE, S. SAMSONOV. *Docking software performance in protein-glycosaminoglycan systems*, in "Journal of Molecular Graphics and Modelling", April 2019, vol. 90, pp. 42-50 [*DOI :* 10.1016/J.JMGM.2019.04.001], https://hal.archives-ouvertes.fr/hal-02391852

[23] N. ZHOU, Y. JIANG, T. BERGQUIST, A. LEE, B. KACSOH, A. CROCKER, K. LEWIS, G. GEORGHIOU, H. NGUYEN, M. N. HAMID, L. DAVIS, T. DOGAN, V. ATALAY, A. RIFAIOGLU, A. DALKIRAN, R. CETIN ATALAY, C. ZHANG, R. HURTO, P. FREDDOLINO, Y. ZHANG, P. BHAT, F. SUPEK, J. FERNANDEZ, B. GEMOVIC, V. PEROVIC, R. DAVIDOVIĆ, N. SUMONJA, N. VELJKOVIC, E. ASGARI, M. R. MOFRAD, G. PROFITI, C. SAVOJARDO, P. L. MARTELLI, R. CASADIO, F. BOECKER, H. SCHOOF, I. KAHANDA, N. THURLBY, A. C. MCHARDY, A. RENAUX, R. SAIDI, J. GOUGH, A. FREITAS, M. ANTCZAK, F. FABRIS, M. WASS, J. HOU, J. CHENG, Z. WANG, A. ROMERO, A. PACCANARO, H. YANG, T. GOLDBERG, C. ZHAO, L. HOLM, P. TÖRÖNEN, A. J. MEDLAR, E. ZOSA, I. BORUKHOV, I. NOVIKOV, A. WILKINS, O. LICHTARGE, P.-H. CHI, W.-C. TSENG, M. LINIAL, P. ROSE, C. DESSIMOZ, V. VIDULIN, S. DZEROSKI, I. SILLITOE, S. DAS, J. G. LEES, D. JONES, C. WAN, D. COZZETTO, R. FA, M. TORRES, A. WARWICK VESZTROCY, J. M. RODRIGUEZ, M. TRESS, M. FRASCA, M. NOTARO, G. GROSSI, A. PETRINI, M. RE, G. VALENTINI, M. MESITI, D. S. ROCHE, J. REEB, D. RITCHIE, S. ARIDHI, S. Z. ALBORZI, M.-D. DEVIGNES, D. C. E. KOO, R. BONNEAU, V. GLIGORIJEVIĆ, M. BAROT, H. FANG, S. TOPPO, E. LAVEZZO, M. FALDA, M. BERSELLI, S. C. TOSATTO, M. CARRARO, D. PIOVESAN, H. UR REHMAN,

Q. MAO, S. ZHANG, S. VUCETIC, G. BLACK, D. JO, E. SUH, J. DAYTON, D. LARSEN, A. OMDAHL, L. MCGUFFIN, D. BRACKENRIDGE, P. BABBITT, J. YUNES, P. FONTANA, F. ZHANG, S. ZHU, R. YOU, Z. ZHANG, S. DAI, S. YAO, W. TIAN, R. CAO, C. CHANDLER, M. AMEZOLA, D. JOHNSON, J.-M. CHANG, W.-H. LIAO, Y.-W. LIU, S. PASCARELLI, Y. FRANK, R. HOEHNDORF, M. KULMANOV, I. BOUDELLIOUA, G. POLITANO, S. DI CARLO, A. BENSO, K. HAKALA, F. GINTER, F. MEHRYARY, S. KAEWPHAN, J. BJÖRNE, H. MOEN, M. E. TOLVANEN, T. SALAKOSKI, D. KIHARA, A. JAIN, T. ŠMUC, A. M. ALTENHOFF, A. BEN-HUR, B. ROST, S. BRENNER, C. ORENGO, C. JEFFERY, G. BOSCO, D. HOGAN, M. MARTIN, C. O'DONOVAN, S. MOONEY, C. GREENE, P. RADIVOJAC, I. FRIEDBERG. *The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens*, in "Genome Biology", December 2019, vol. 20, n^{o} 1 [*DOI :* 10.1186/s13059-019-1835-8], https://hal.inria.fr/hal-02393202

[24] M. ZOGHLAMI, S. ARIDHI, M. MADDOURI, E. MEPHU NGUIFO. *Multiple instance learning for sequence data with across bag dependencies*, in "International journal of machine learning and cybernetics", 2019, https://arxiv.org/abs/1602.00163 [*DOI :* 10.1007/s13042-019-01021-5], https://hal.inria.fr/hal-02393742

International Conferences with Proceedings

[25] B. SARKER, D. W. RITCHIE, S. ARIDHI. *Functional Annotation of Proteins using Domain Embedding based Sequence Classification*, in "KDIR 2019 - 11th International Conference on Knowledge Discovery and Information Retrieval", Vienna, Austria, September 2019, https://hal.inria.fr/hal-02283430

[26] M. ZOGHLAMI, S. ARIDHI, M. MADDOURI, E. M. NGUIFO. *A Structure Based Multiple Instance Learning Approach for Bacterial Ionizing Radiation Resistance Prediction*, in "KES 2019 - 23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems", Budapest, Hungary, September 2019, https://hal.inria.fr/hal-02307048

National Conferences with Proceedings

[27] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Les forêts d'arbres extrêmement aléatoires : utilisation dans un cadre non supervisé*, in "EGC 2019 - 19ème Conférence Francophone sur l'Extraction et Gestion des connaissances", Metz, France, Hermann-Éditions, January 2019, vol. RNTI E-35, pp. 395-400, https://hal.inria.fr/hal-02099532

Conferences without Proceedings

[28] M.-E. RUIZ-ECHARTEA, I. CHAUVOT DE BEAUCHÊNE, D. RITCHIE. *EROS-DOCK for Pairwise and Multi-body Protein-Protein Docking*, in "Journée MASIM2019 (Méthodes Algorithmiques pour les Structures et Interactions Macromoléculaires)", Paris, France, November 2019, https://hal.archives-ouvertes.fr/hal-02391973

[29] A. VAGINAY, M. SMAIL-TABBONE, T. BOUKHOBZA. *Towards an automatic conversion from SBML core to SBML qual*, in "Journées Ouvertes Biologie, Informatique et Mathématiques, JOBIM 2019", Nantes, France, July 2019, Présentation Poster, https://hal.archives-ouvertes.fr/hal-02407443

Other Publications

[30] F.-É. CALVIER, P. MONNIN, M. BOLAND, P. JARNOT, E. BRESSO, M. SMAÏL-TABBONE, A. COULET, C. BOUSQUET. *Providing Molecular Characterization for Unexplained Adverse Drug Reactions : Podium Abstract*, July 2019, Podium Abstract at MedInfo 2019, Lyon, France, https://hal.inria.fr/hal-02196134

[31] K. DALLEAU, M. COUCEIRO, M. SMAIL-TABBONE. *Computing Vertex-Vertex Dissimilarities Using Random Trees: Application to Clustering in Graphs*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02427563

[32] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Clustering graphs using random trees*, September 2019, working paper or preprint, https://hal.inria.fr/hal-02282207

[33] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Unsupervised Extra Trees: a stochastic approach to compute similarities in heterogeneous data.*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01982232

[34] W. INOUBLI, S. ARIDHI, H. MEZNI, M. MONDHER, E. M. NGUIFO. *A Distributed Algorithm for Large-Scale Graph Clustering*, August 2019, working paper or preprint, https://hal.inria.fr/hal-02190913

[35] M.-E. RUIZ-ECHARTEA, I. CHAUVOT DE BEAUCHÊNE, D. RITCHIE. *EROS: A Protein Docking Algorithm Using a Quaternion pi- Ball Representation for Exhaustive and Accelerated Exploration of 3D Rotational Space*, April 2019, GGMM (groupe de graphisme et modélisation moléculaire), Poster, https://hal.archives-ouvertes.fr/hal-02392106

References in notes

[36] C. E. ALVAREZ-MARTINEZ, P. J. CHRISTIE. *Biological diversity of prokaryotic type IV secretion systems*, in "Microbiology and Molecular Biology Reviews", 2011, vol. 73, pp. 775–808

[37] M. BAADEN, S. R. MARRINK. *Coarse-grained modelling of protein-protein interactions*, in "Current Opinion in Structural Biology", 2013, vol. 23, pp. 878–886

[38] A. BERCHANSKI, M. EISENSTEIN. *Construction of molecular assemblies via docking: modeling of tetramers with D_2 symmetry*, in "Proteins", 2003, vol. 53, pp. 817–829

[39] P. BORK, L. J. JENSEN, C. VON MERING, A. K. RAMANI, I. LEE, E. M. MARCOTTE. *Protein interaction networks from yeast to human*, in "Current Opinion in Structural Biology", 2004, vol. 14, pp. 292–299

[40] I. J. CHAUVOT DE BEAUCHENE, S. J. DE VRIES, M. J. ZACHARIAS. *Fragment-based modeling of protein-bound ssRNA*, September 2016, ECCB 2016: The 15th European Conference on Computational Biology, Poster, https://hal.archives-ouvertes.fr/hal-01573352

[41] I. CHAUVOT DE BEAUCHÊNE, S. J. DE VRIES, 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], https://hal.archives-ouvertes.fr/hal-01505862

[42] S. J. DE VRIES, I. CHAUVOT DE BEAUCHÊNE, C. E. M. SCHINDLER, M. ZACHARIAS. *Cryo-EM Data Are Superior to Contact and Interface Information in Integrative Modeling*, in "Biophysical Journal", February 2016 [*DOI :* 10.1016/J.BPJ.2015.12.038], https://hal.archives-ouvertes.fr/hal-01505863

[43] M.-D. DEVIGNES, S. BENABDERRAHMANE, M. SMAÏL-TABBONE, A. NAPOLI, 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”", 2012, vol. 5, n^{o} 3/4, pp. 245-260, https://hal.inria.fr/hal-00734329

[44] S. E. DOBBINS, V. I. LESK, 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", 2008, vol. 105, n^{o} 30, pp. 10390–10395

[45] M. EL HOUASLI, B. MAIGRET, M.-D. DEVIGNES, A. W. GHOORAH, S. GRUDININ, D. RITCHIE. *Modeling and minimizing CAPRI round 30 symmetrical protein complexes from CASP-11 structural models*, in "Proteins: Structure, Function, and Genetics", March 2017, vol. 85, n^{o} 3, pp. 463–469 [*DOI :* 10.1002/PROT.25182], https://hal.inria.fr/hal-01388654

[46] W. J. FRAWLEY, G. PIATETSKY-SHAPIRO, C. J. MATHEUS. *Knowledge Discovery in Databases: An Overview*, in "AI Magazine", 1992, vol. 13, pp. 57–70

[47] R. FRONZES, E. SCHÄFER, L. WANG, H. R. SAIBIL, E. V. ORLOVA, G. WAKSMAN. *Structure of a type IV secretion system core complex*, in "Science", 2011, vol. 323, pp. 266–268

[48] R. A. GOLDSTEIN. *The structure of protein evolution and the evolution of proteins structure*, in "Current Opinion in Structural Biology", 2008, vol. 18, pp. 170–177

[49] 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, R. APWEILER. *The HUPO PSI's Molecular Interaction format – a community standard for the representation of protein interaction data*, in "Nature Biotechnology", 2004, vol. 22, n^{o} 2, pp. 177-183

[50] H. I. INGÓLFSSON, C. A. LOPEZ, J. J. UUSITALO, D. H. DE JONG, S. M. GOPAL, X. PERIOLE, S. R. MARRINK. *The power of coarse graining in biomolecular simulations*, in "WIRES Comput. Mol. Sci.", 2013, vol. 4, pp. 225–248, http://dx.doi.org/10.1002/wcms.1169

[51] J. D. JACKSON. *Classical Electrodynamics*, Wiley, New York, 1975

[52] P. J. KUNDROTAS, Z. W. ZHU, I. A. VAKSER. *GWIDD: Genome-wide protein docking database*, in "Nucleic Acids Research", 2010, vol. 38, pp. D513–D517

[53] M. F. LENSINK, S. J. WODAK. *Docking and scoring protein interactions: CAPRI 2009*, in "Proteins", 2010, vol. 78, pp. 3073–3084

[54] A. MAY, M. ZACHARIAS. *Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking*, in "Proteins", 2008, vol. 70, pp. 794–809

[55] I. H. MOAL, P. A. BATES. *SwarmDock and the Use of Normal Modes in Protein-Protein Docking*, in "International Journal of Molecular Sciences", 2010, vol. 11, n^{o} 10, pp. 3623–3648

[56] C. MORRIS. *Towards a structural biology work bench*, in "Acta Crystallographica", 2013, vol. PD69, pp. 681–682

[57] D. MUSTARD, D. RITCHIE. *Docking essential dynamics eigenstructures*, in "Proteins: Structure, Function, and Genetics", 2005, vol. 60, pp. 269-274 [*DOI :* 10.1002/PROT.20569], https://hal.inria.fr/inria-00434271

[58] 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, H. HERMJAKOB. *Protein interaction data curation: the International Molecular Exchange (IMEx) consortium*, in "Nature Methods", 2012, vol. 9, n^{o} 4, pp. 345-350

[59] B. PIERCE, W. TONG, Z. WENG. *M-ZDOCK: A Grid-Based Approach for C_n Symmetric Multimer Docking*, in "Bioinformatics", 2005, vol. 21, n^{o} 8, pp. 1472–1478

[60] D. RITCHIE, G. J. KEMP. *Protein docking using spherical polar Fourier correlations*, in "Proteins: Structure, Function, and Genetics", 2000, vol. 39, pp. 178-194, https://hal.inria.fr/inria-00434273

[61] D. RITCHIE, D. KOZAKOV, S. VAJDA. *Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational FFT generating functions*, in "Bioinformatics", June 2008, vol. 24, n^{o} 17, pp. 1865-1873 [*DOI :* 10.1093/BIOINFORMATICS/BTN334], https://hal.inria.fr/inria-00434264

[62] D. RITCHIE. *Recent Progress and Future Directions in Protein-Protein Docking*, in "Current Protein and Peptide Science", February 2008, vol. 9, n^{o} 1, pp. 1-15 [*DOI :* 10.2174/138920308783565741], https://hal.inria.fr/inria-00434268

[63] A. RIVERA-CALZADA, R. FRONZES, C. G. SAVVA, V. CHANDRAN, P. W. LIAN, T. LAEREMANS, E. PARDON, J. STEYAERT, H. REMAUT, G. WAKSMAN, E. V. ORLOVA. *Structure of a bacterial type IV secretion core complex at subnanometre resolution*, in "EMBO Journal", 2013, vol. 32, pp. 1195–1204

[64] M. E. RUIZ ECHARTEA, I. CHAUVOT DE BEAUCHÊNE, D. RITCHIE. *Accelerating Protein Docking Calculations using the ATTRACT CoarseGrained Force Field and 3D Rotation Maps*, May 2017, GGMM-2017, Poster, https://hal.inria.fr/hal-01927271

[65] M. G. SAUNDERS, G. A. VOTH. *Coarse-graining of multiprotein assemblies*, in "Current Opinion in Structural Biology", 2012, vol. 22, pp. 144–150

[66] D. SCHNEIDMAN-DUHOVNY, Y. INBAR, R. NUSSINOV, H. J. WOLFSON. *Geometry-based flexible and symmetric protein docking*, in "Proteins", 2005, vol. 60, n^{o} 2, pp. 224–231

[67] M. L. SIERK, G. J. KLEYWEGT. *Déjà vu all over again: Finding and analyzing protein structure similarities*, in "Structure", 2004, vol. 12, pp. 2103–2011

[68] V. VENKATRAMAN, D. RITCHIE. *Flexible protein docking refinement using pose-dependent normal mode analysis*, in "Proteins", June 2012, vol. 80, n^{o} 9, pp. 2262-2274 [*DOI :* 10.1002/PROT.24115], https://hal.inria.fr/hal-00756809

[69] A. B. WARD, A. SALI, I. A. WILSON. *Integrative Structural Biology*, in "Biochemistry", 2013, vol. 6122, pp. 913–915

[70] 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, B. HONIG. *Structure-based prediction of protein-protein interactions on a genome-wide scale*, in "Nature", 2012, vol. 490, pp. 556–560

[71] A. ÖZGUR, Z. XIANG, D. R. RADEV, Y. HE. *Mining of vaccine-associated IFN-γ gene interaction networks using the Vaccine Ontology*, in "Journal of Biomedical Semantics", 2011, vol. 2 (Suppl 2), S8 p.