RESEARCH CENTRE

Nancy - Grand Est

IN PARTNERSHIP WITH:

Université de Lorraine, CNRS

2021 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 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]
- Isaure Chauvot de Beauchêne [CNRS, Researcher]
- Kévin Dalleau [Univ de Lorraine, Researcher, until Aug 2021]
- Bernard Maigret [CNRS, Emeritus]
- Athénaïs Vaginay [Univ de Lorraine, Researcher, from Oct 2021]

Faculty Members

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

Post-Doctoral Fellows

- Camille Besançon [Inria, until Feb 2021]
- Dominique Mias-Lucquin [Univ de Lorraine]

PhD Students

- Diego Amaya Ramirez [Inria]
- Hrishikesh Dhondge [CNRS]
- Kamrul Islam [Univ de Lorraine]
- Anna Kravchenko [CNRS]
- Antoine Moniot [Univ de Lorraine, until Sep 2021]
- Bishnu Sarker [Inria, until Feb 2021]
- Athénaïs Vaginay [Centre de recherche en automatique de Nancy, until Sep 2021]

Technical Staff

- Antoine Moniot [CNRS, Engineer, from Oct 2021]
- Bishnu Sarker [CNRS, Engineer, from Mar 2021 until May 2021]
- Louane Sigrist [Univ de Lorraine, Engineer, until Oct 2021]

Interns and Apprentices

- Nicolas Bombarde [Univ de Lorraine, from Jun 2021 until Jul 2021]
- Ugo Cottin [Univ de Lorraine, from Jun 2021 until Jul 2021]
- Alix Delannoy [Univ de Lorraine, until Feb 2021]
- Karina Pats [Agence Erasmus+ France, from Feb 2021 until Jun 2021]
- Anna Perez Rafols [Giotto Biotech (It), Jun 2021]
- Valentin Retter [CNRS, from May 2021 until Jul 2021]

- Dalil Rouabah [Univ de Lorraine, from Apr 2021 until Sep 2021]
- Amal Stiti [Univ de Lorraine, from Feb 2021 until Apr 2021]

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 [67, 84].

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 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 [59, 79]. 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 [62, 85]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [49].

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 [44, 45, 80]. In particular it will be interesting to set up fuzzy clustering methods taking advantage of our previous work on gene functional classification [54], 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) [57].

For example, domain family classification [48, 65] is relevant for studying domain-domain interactions (DDI). Our previous work on Knowledge-Based Docking (KBDOCK, [4, 6]) 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 [69]. 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 Enzyme Commission (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 constituted Bishnu Sarker's PhD thesis (defended in April 2021) [37] with a major contribution called GrAPFI (Graph-based Automatic Protein Function Inference) and is now continued in Md Kamrul Islam's thesis (ongoing since 2019).

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 [72], 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 [61]

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

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 FFT techniques may be used to accelerate the search in up to five of the six degrees of freedom [73]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [8, 11]. 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, 71], 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 [11], [73]. 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 [47, 70, 78]. 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 [12], and was subsequently applied to several symmetrical complexes from the "CAPRI" blind docking experiment [56]. 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 [55, 64, 66, 68]. In our experience, docking ensembles of NMA conformations do not give much improvement over basic FFT-based soft docking [83], 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 [46]. 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 [77]. Furthermore, this kind of CG model effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [60]. 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 [53], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space, as intiated with the EROS-DOCK software during the PhD project of Maria Elisa Ruiz Echartea (defended in 2019)[13, 14].

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 [50, 51].

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 have recently developped a new representation for the clustering of 3D conformations based on internal coordinates, in order to optimise the representativity of the library (internship of Alix Delannoy). We are currently developping a new clustering method based on epsilon-networks to reduce the cardinality of the ensemble of clusters.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, Kevin Dalleau, Bishnu Sarker, Kamrul Islam, 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 [82]. Our approaches will be driven by knowledge and will keep close connection with experimental data.

4.2 Prokaryotic Type IV Secretion Systems

Participants: Isaure Chauvot de Beauchêne *(contact person)*, Marie-Dominique Devignes, Bernard Maigret, 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 [43]. 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 [58, 74]. 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 [81]. 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.

4.4 3D structual differences among HLA antigens

Participants: Marie-Dominique Devignes *(contact person)*, Diego Amaya Ramirez, Louane Sigrist, Bernard Maigret.

A third application domain has emerged in Axis 2 through the Inria-Inserm PhD thesis of Diego Amaya Ramirez, in collaboration with the Immunology and Histocompatibility Laboratory at the APHP Saint-Louis Hospital in Paris. Differences between donor and recipient HLA proteins are one of the major limitations of organ transplant because of HLA ubiquity on cells of tissues and organs. Indeed, in case of incompatibility between the HLA proteins of the donor and those of the patient, an immune response is triggered in the patient that can result in rejection of the transplanted organ. The thesis aims at deciphering the role played by tiny 3D structure differences between donor and recipient HLA proteins in determining the production of donor-specific antiodies by the recipient. We are currently developing methods to compare locally structure variations between HLA proteins, taking into account the dynamics of these proteins.

5 Social and environmental responsibility

5.1 Environmental Footprint of Research Activities

Many conferences have been run online in 2021. This has lead to a strong decrease in the environmental footprint of the team with respect to plane travels. In particular Athénaïs Vaginay arranged a local event for attending together but remotely the JOBIM conference (held in Paris, 6-9 july 2021). Several PhD students presented their work online at international conferences in 2021.

6 Highlights of the year

6.1 Awards

Dominique Mias-Lucquin obtained the best poster award at the joint 22th *GGMM (Group of Graphism and Molecular Modeling)* and 10th *SFCi (French Society of Chemoinformatics)* meetings (Lille, 29 Sept - 1 Oct 2021).

7 New software and platforms

7.1 New software

7.1.1 HLA-3D-Diff

Name: Interface for viewing and superimposing 3D structures of HLA antigens

Keywords: Visualization, 3D structure, Proteins, Molecular dynamics, GUI (Graphical User Interface)

Scientific Description: The HLA-3D-Diff software contains two parts: (1) The HLA-3D-Diff database: its goal is to provide easy access to data from multiple public sources and data generated by members of the project , (2) The HLA-3D-Diff visualisation interface: a web app for visualizing structures and molecular dynamics (MD) simulations of HLA molecules.

The GitLab repository contains: - The source code and Docker environment for these tools , - Documentation for the database, including models, the data dictionary, and input data , - Documentation for the interface (a short video).

Functional Description: This tool makes it possible to visually search for 3D structural differences that may explain incompatibility or unexpected compatibility between two distinct HLA antigens. It is differential because it makes it possible to superimpose two different 3D structures, and dynamic because it makes it possible to visualize trajectories simulated by molecular dynamics.

URL: https://gitlab.inria.fr/capsid/hla-3d-diff

Contact: Marie-Dominique Devignes

Participants: Louane Sigrist, Diego Amaya Ramirez, Marie-Dominique Devignes, Jean-Luc Taupin

7.2 New platforms

Participants: Marie-Dominique Devignes *(scientific responsible)*, Malika Smaïl-Tabbone *(contact person)*, Sabeur Aridhi, Bernard Maigret, Antoine Moniot, Diego Amaya Ramirez.

The CAPSID team is actively involved with the Orpailleur Team in the evolution of the MBI-DS4H research platform for an increased sharing of resources in structural bioinformatics (MBI: Modeling Biomolecules and their Interactions) and data science for health (DS4H).

Thanks to the Contrat de Plan Etat-Region (CPER IT2MP), new equipment has been acquired and installed in concertation with the Grid5K project. NAMD3 version has been installed on specific GPUs to accelerate computation time for molecular dynamics. Benchmarking datasets for clustering of mixed data are offered to download in the Data Science section of the platform.

The technical support of the patform is ensured by the LORIA SISR (Service d'Ingénierie en Soutien de la Recherche) and can be followed on gitlab.

8 New results

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

Participants: Marie-Dominique Devignes, Malika Smaïl-Tabbone, Sabeur Aridhi, Kevin Dalleau, Bishnu Sarker, Kamrul Islam, Athénaïs Vaginay.

8.1.1 Biomedical Knowledge Discovery

We updated and extended a preliminary study carried out by Zia Alborzi (Inria PhD student 2015-2018) during his PhD thesis and aimed at inferring protein domain interactions from multiples sources of protein-protein interactions [1, 34]. Results have been deposited as open research data for reuse facility in the zenodo repository.

We continued our collaboration with clinicians at the CHRU Nancy in the frame of the RHU FIGHT-HF program. In particular, we contributed to a study aimed at clustering patients according to their echocardiographic measurements and at characterizing the obtained clusters with respect to heart failure propensity [23].

In the frame of the PraktikPharma ANR project coordinated by Adrien Coulet (Orpaileur team), we investigated adverse drug reaction mechanisms with knowledge graph mining [20]. Finally we also participated in a study carried out in the Orpailleur team in the frame of the GeenAge IMPACT project to tackle the problem of identifying the most predictive and discriminant features in supervised classification problems [29].

8.1.2 Stochastic Random Trees for Similarity Computation

Kevin Dalleau defended his PhD thesis entitled "A stochastic approach based on random forests for dissimilarity calculation : application to clustering for various data structured" on 23 november 2021 [52]. The UET dissimilarity measure introduced in this thesis was used in a benchmarking study of clustering algorithms applied on mixed data (both categorical and numerical) [9].

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 [76]. 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 [75].

A preliminary exploration study of the potential of generative adversarial networks (GAN) for protein function annotation was conducted by Bishnu Sarker. The proposed method “Prot-A-GAN” uses GAN-like adversarial training for learning embedding of nodes and relation in a heterogeneous knowledge graph [32]. Prot-A-GAN trains a discriminator using domain-adaptive negative sampling to discriminate positive and negative triples, and then, it trains a generator to guide a random walk over the knowledge graph that identify paths between proteins and GO annotations.

All these contributions constitute the PhD Thesis of Bishnu Sarker whose defense took place on April 23, 2021 [37].

In the context of the “TempoGraphs” project, we developed a similarity function for sparse binary data with application on protein function annotation [26]. The proposed similarity function is based on the analysis of the best existing similarity functions for the protein function annotation task. We performed experiments in a simple pairwise similarity scenario and also using our proposal as part of a more complex protein function annotation method.

In the frame of Kamrul Islam’s PhD project, we first tackled the problem of link prediction in large knowledge graphs. We conducted an experimental evaluation of similarity-based and embedding-based link prediction methods on graphs. The studied approaches were evaluated on several graph datasets with different properties from various domains. The precision of similarity-based approaches was computed in two different ways to highlight the difficulty of tuning the threshold for deciding the link existence based on the similarity score [22].

Then, we tackled the problem of negative sampling, in other words: how to sample negative triplets from non-observed ones in a training knowledge graph. We developed a simple negative sampling (SNS) method based on the assumption that the entities which are closer to the corrupted entity in the embedding space are able to provide high-quality negative triples. The proposed method has been evaluated through link prediction task on several popular knowledge graph datasets including a new biological knowledge graph dataset (FIGHT-HF-23R) [35].

Collaborative work on the development of a distributed and incremental algorithm for large-scale graph clustering has led to a technical report in 2020 [38] which contributed to the completion of Wissem Inoubli’s PhD thesis (defended early January 2021) [36].

Another collaborative work on the study of big services has led to a review paper in 2021 [24]. The purpose of the study is to provide an understanding of the new emerging big service model from the point of view of "lifecycle management phases". We also study the role of big data frameworks in the provisioning of big services [24].

8.1.4 Biological network modelling

An inter-disciplinary project, funded by the Université de Lorraine, is ongoing since 2018 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 [21].

Boolean Networks (BNs) are a simple formalism used to study complex biological systems when the prediction of exact reaction times is not of interest. BNs play a key role in understanding the dynamics of the studied systems and in predicting their disruption in case of complex human diseases. BNs are generally built from experimental data and knowledge from the literature, either manually or with the aid of programs. The automatic synthesis of BNs is still a challenge for which several approaches have been proposed. We proposed ASKeD-BN, a new approach based on Answer-Set Programming to synthesise BNs constrained in their structure and dynamics [27]. By applying ASKeD-BN on several well-known biological systems, we provide empirical evidence that our approach can construct BNs in line with the provided constraints. We compared our approach with three existing methods (REVEAL, Best-Fit and caspo-TS) and showed that our approach produces a small number of BNs which are covering a good

proportion of the dynamical constraints, and that the variance of this coverage is low. The published paper on ASKed-BN was selected as a highlight oral presentation in the JOBIM 2021 conference [33].

Today SBML is the standard format to represent models of biological systems. Most of the established curated models available in the Biomodels repository are quantitative, but in some cases qualitative models, such as boolean networks are preferable. We proposed to focus on the automatic transformation of quantitative SBML models to boolean networks through the so-called SBML2BN pipeline. By running SBML2BN on more than 200 quantitative SBML models, we provided evidence that we can automatically construct Boolean networks which are compatible with both the structure and the dynamics of a given quantitative SBML model [28, 34].

8.2 Axis 2 : Integrative Multi-Component Assembly and Modeling

Participants: Isaure Chauvot de Beauchêne, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Bernard Maigret, Dominique Mias-Lucquin, Antoine Moniot, Anna Kravchenko, Hrishikesh Dhondge, Alix Delannoy, Diego Amaya Ramirez, Louane Sigrist.

8.2.1 Clustering of 3D conformations

Our fragment-based approach to dock RNA on proteins requires libraries of 3D conformations of RNA fragments. These must be representative of the full ensemble of possible 3D conformations of RNA fragments but be of small enough cardinality to be usable for combinatorial assemblies of fragments into RNAs. Therefore, they must approximate all possible local RNA conformations within a given precision, by a set of well chosen representative fragments. Such a set can be obtained by clustering a larger set of fragments (typically 10^4 to 10^5) whose structures have been solved experimentally, using suitable clustering algorithm and measure of dissimilarity between fragments. A commonly used measure of dissimilarity in structural biology is the root mean square deviation (RMSD), whose exact computation requires a pairwise structural alignment. But this alignment is highly time-consuming and not applicable for a very large initial set of fragments. We have developed an approach based on feature extraction to perform an effective clustering, while avoiding a computationally expensive full pairwise alignment. Using as example poly-A RNA fragments of 3 nucleotides (3-nt), we searched for internal coordinates whose differences can best approximate the RMSD between two fragments without any superposition. We found that the simple differences of internal distances and angles can provide a lower bound on the RMSD, allowing us to filter out redundant fragment pairs without any RMSD computing [30]. This was developed by Alix Delannoy (3A Mines research training), co-supervised by Antoine Moniot, Isaure Chauvot de Beauchêne and Yann Guermeur. Concerning the clustering step, we are currently working on an algorithm to compute epsilon-nets of small cardinality for finite sets of points. Each of these points should be at a distance inferior to epsilon to a point in the net. The minimal cardinality that can be achieved is called the (epsilon) covering number. Although the literature provides us with upper bounds on the covering numbers, the so-called extended Sauer-Shelah lemmas, the proofs are non constructive. To the best of our knowledge, there are no (efficient) algorithms to compute epsilon-nets of small cardinality in the framework of interest. We have chosen to develop an algorithm, based on two ideas: the ascending hierarchical clustering (AHC) and the minimal enclosing spheres. The hierarchical approach makes it possible to go from a feasible solution to another and stop as soon as no more fusion is possible [41, 42]. This work is carried out by Antoine Moniot for his PhD thesis, to be defended in 2022.

8.2.2 Modeling and design of RNA-protein complexes

Our H2020 ITN project RNAct aims at designing new RNA-binding proteins based on the well-conserved protein domain called RNA Recognition Motif (RRM). In this frame, we have created in 2020 the InteR3Mdb database that contains all known 3D structures of RRM and RRM-RNA complexes. We have now superimposed all those structures and aligned their sequences (which is not trivial due to sequence and structure variability, large deletions, inconsistencies in the raw data...), and identified all nucleotide – amino acid contacts in each experimental structure of an RRM-RNA complex. For each contact, the

amino acid type and position in the sequence, the nucleotide type and position in the sequence, the type of interaction (hydrogen bond, ionic bond, stacking...), the atoms involved and the atom-atom distances were identified. This information is extremely useful and provides an important resource on its own to the scientific community. Importantly, it allowed us to infer a contact propensity for each amino acid at every conserved position in the sequence, and toward a given nucleotide type (A/C/U/G). This is currently used to computationally design and propose point mutations to alter RNA binding specificity. Synthesis of RRM proteins carrying such mutations is underway within the RNAct Consortium in view of performing binding assays that can validate the predictions.

In parallel, we exploited the InteR3Mdb database to tackle the bottleneck encoutered when modeling RNA-protein complexes using the known 3D shape of the unbound protein. Indeed, the changes that occur in the protein 3D shape when it binds RNA are hard to predict. In InteR3Mdb (Hrishikesh Dhondge's PhD thesis), the protein-bound or protein-unbound status of all known RRM 3D structures have been added as new attributes for each database entry. Then, data was processed to infer those structural characteristics that could distinguish bound from unbound RRMs. Results can be used for the prediction of the bound conformation of an RRM from its unbound conformation. The knowledge and prediction power gained from the bound/unbound comparison of RRMs will enable us to (i) unravel structural determinants of the RNA binding capacities of RRMs, (ii) address the main limitation of RRM – RNA docking approaches, and (iii) take into account conformational changes associated with RNA binding in the design of RRMs with altered RNA binding specificity.

In collaboration with Sjoerd de Vries (INSERM), we have implemented a new method to create energy parameters for RNA-protein interactions in coarse-grained representations. Each amino-acid of the protein and each nucleotide of the RNA is represented by 2 to 7 pseudo-atoms ("beads"). For each model of an RNA - protein interaction, the energy is computed as the sum of the bead-bead energies, and the model with the lowest energy is considered as the most probable. Each bead-bead energy depends solely on the 2 bead types (among 17 RNA types and 32 protein types) and the inter-bead distance. For each pair of bead types, the energy function has the shape of a Lennard-Jones potential, given by 2 parameters that determine the distance of minimal energy and the value of the minimal energy. The current parameters were extracted in 2010 by statistics on existing RNA-protein crystal structures and optimized by a random Monte Carlo-like strategy. These parameters were not initially tailored to single-stranded RNA and their performance is not flawless. A main goal of Anna Kravchenko's thesis is to optimize those parameters for ssRNA. To achieve that, we set up a novel histogram-based approach. For each pair of bead types, we (a) convert the current energy function into a log-odds histogram of the expected occurrences of distances (discretized into bins) in correct/incorrect models (from an in-house benchmark), using the Boltzmann equation; (b) obtain the corresponding histogram on a benchmark-wide docking test, which corresponds to the residual error of the energy function; (c) sum the predicted and real histograms; (d) analytically fit the energy parameters to the resulting histogram; repeat until convergence, i.e. until the residual histogram is flat. A first prototype of this workflow has been validated on a small training set [40], and is now being extended to our full benchmark of hundred of protein-RNA complexes.

8.2.3 3D Modeling and Virtual Screening

3D modeling has been revolutionized by the AlphaFold2 program created by DeepMind using deep learning methods. We used the open access AlphaFold2 program to confirme and refine our models of the Relaxase RelSt3 dimer in the frame of the CITRAM project. The resulting model was used to simulate the interaction with both ss and ds RNA. This work allowed Dominique Mias-Lucquin to win the best Poster award at GGMM'2021 (An homodimeric multi-domain protein interacting with ssDNA and dsDNA: The challenges of RelSt3 relaxase modelisation).

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. Our work on the search of inhibitors of beta1,4 galactosyltransferase 7 was presented at the RARE meeting [39].

9 Bilateral contracts and grants with industry

The CAPSID team has no bilateral contracts and grants with industry.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria associate team not involved in an IIL or an international program

FlexMol

Participants: Marie-Dominique Devignes, Isaure Chauvot de Beauchêne, Antoine Moniot, Dominique Mias-Lucquin.

Title: Algorithms for Multiscale Macromolecular Flexibility

Duration: 2019 - 2022

Principal investigator (Inria): Sergei Grudinin, Inria Nano-D until 2020, then DAO team (DATA department) at the Laboratoire Jean Küntzmann, UMR 5224 (CNRS, Grenoble INP, Inria, UGA), Grenoble.

Principal investigator (partner institution): Pablo Chacon, Rocasolano Institute of Physical Chemistry (IQFR-CSIC), Madrid, Spain.

Inria contact: Marie-Dominique Devignes

Summary: Molecular flexibility is essential to link structure and function of many biological macromolecules and one of the main challenges in the field of computational structural biology is to predict and explain molecular flexibility and corresponding conformational changes. The main goal of this collaboration is to mutually compare, combine and explore possible applications of novel computational techniques for emerging problems in structural biology and bioinformatics related to molecular flexibility.

10.1.2 Participation in other International Programs

Tempographs: The Tempograph project was prolongated till the end of 2021 due to the Covid crisis.

Participants: Sabeur Aridhi, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Bishnu Sarker.

Title: Analysing big data with temporal graphs and machine learning

Duration: 2019 - 2021

Coordinator: Sabeur Aridhi (sabeur.aridhi@loria.fr)

Partners:

- LIMOS, Université de Clermont Auvergne, France.
- DAVID, Université de Versailles Saint-Quentin, France.
- Universidade Federale do Ceara, Fortaleza, Brazil.

Inria contact: Sabeur Aridhi

Summary: In the TempoGraphs project, we investigate and propose solutions, based on temporal graphs and machine learning, for both urban traffic-related problems and protein annotation problems.

10.2 European initiatives

10.2.1 FP7 & H2020 projects

H2020 ITN RNAct

Participants: Isaure Chauvot de Beauchêne, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Anna Kravchenko, Hrishikesh Dhondge, Antoine Moniot.

Title: Innovative Training Network: Enabling proteins with RNA recognition motifs for synthetic biology and bio-analytics.

Duration: October 2018 - October 2022

Coordinator: Wim Vranken (Vrije University Bruxelles, Belgium)

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)

Inria contact: Isaure Chauvot de Beauchêne

Summary: The RNAct 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.

10.3 National initiatives

3D structural differences among HLA proteins

Participants: Marie-Dominique Devignes *(contact person)*, Diego Amaya Ramirez, Louane Sigrist, Bernard Maigret.

Context : PhD thesis Inria-Inserm of Diego Amaya Ramirez (2019-2022). This thesis (HLA genetic system and organ transplantation: understanding the basics of immunogenicity to improve donor - receptor compatibility when assigning grafts to recipients) is co-supervised by Marie-Dominique Devignes and Jean-Luc Taupin (Professor of Immunology at the Laboratory of Immunology and Histocompatibility at Saint-Louis Hospital and Diderot University in Paris).

Predicting the 2D structure of protein-bound RNAs

Participants: Isaure Chauvot de Beauchêne.

Context: Co-supervision of a M2 internship with Fariza Tahi and Eric Angel (AROBAS team) at the IBISC lab (University of Evry). This project consisted in integrating the CAPSID library of protein-bound RNA fragments into the 2D RNA prediction pipeline of the AROBAS team to create specific RNA predictions in the protein-binding context. It provided a proof-of-principle, and a demand for a PhD grant has been submitted to pursue the project in 2022.

FIGHT-HF

Participants: Marie-Dominique Devignes, 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: 2015–2021. 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. 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.

Anti-covid research

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

Our research on the structural molecular determinants of binding of SARS-Cov-2 variant Spike proteins to their cellular receptor has lead us to develop partnerhsips with selected teams in France.

- LAMA : Laboratoire de Mathématique, CNRS — Université Savoie Mont Blanc UMR 5127. Persons involved : Laurent Vuillon, Aria Gheeraert.
- IRIM : Institut de Recherche en Infectiologie de Montpellier, CNRS-Université de Montpellier UMR 9004. Persons involved : Laurent Chaloin, Olivier Montcorgé.

10.4 Regional initiatives

CPER IT2MP Santé (2016-2021) In the regional Contrat de Plan Etat-Région Santé (Innovations Technologiques, Modélisation et Médecine Personnalisée), MD Devignes and Malika Smaïl-Tabbone coordinate the SMEC platform (Simulation, Modélisation, Extraction de Connaissances). This has strenghtened our partnership with two teams at the regional level.

- LPCT (Laboratoire de Physique et Chimie Théoriques, CNRS-Université de Lorraine UMR 7019), Team of Chris Chipot and François Dehez.
- CIC-P (Centre d'Investigation Clinique Plurithématique, CHRU de Nancy) and DCAC (Défaillance Cardiovasculaire Aiguë et Chronique, Inserm - Université de Lorraine U1116), Patrick Rossignol, Nicolas Girerd, Patrick Lacolley and Véronique Régnault.

LUE IMPACT projects (2017-2021) The CAPSID team has been associated with two LUE projects that terminated in 2021

- IMPACT GeenAge. MD Devignes (contact person). Partners involved : IMOPA (Ingénierie Moléculaire et Physiopathologie Articulaire, CNRS - Université de Lorraine, UMR 7365), Isabelle Behm ; Orpailleur (CRI Nancy Grand-Est), Amedeo Napoli, Alexandre Bazin (GeenAge post-doc).
- IMPACT BioMolecules. Sabeur Aridhi (contact person). Partners involved : Laboratoire d'Ingenierie des Biomolecules (ENSAIA - Université de Lorraine), Freédéric Borges ; Orpailleur (CRI Nancy Grand-Est), Yannick Toussaint.

LUE-FEDER CITRAM (2017-2022) The CITRAM project (Conception d'Inhibiteurs de la Transmission des Résistances Anti-Microbiennes) was created as a LUE project in 2017 and later extended with FEDER co-funding. It involves two other teams:

- DynAMIc lab (Genome dynamics and microbial adaptation, INRAE - Université de Lorraine UMR 1128), Team of Nathalie Leblond and Nicolas Soler.
- LPCT (Laboratoire de Physique et Chimie Théoriques, CNRS-Université de Lorraine UMR 7019), Team of Chris Chipot and François Dehez.

11 Dissemination

Participants: Marie-Dominique Devignes, Malika Smaïl-Tabbone, Isaure Chauvot de Beauchêne, Sabeur Aridhi.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- Sabeur Aridhi was member of the organizing committee of the 2021 *ICML Workshop on Computational Biology*.
- Isaure Chauvot de Beauchêne was member of the organizing committee of the *AlgoSB winter school* 2021.

11.1.2 Scientific events: selection

Member of conference program committees

- Isaure Chauvot de Beauchêne was member of the conference program committees of *JOBIM* 2021 and *GGMM* 2021.
- Marie-Dominique Devignes was member of the conference program committee of *IEEE BIBM* 2021.

11.1.3 Journal

Member of the editorial boards Marie-Dominique Devignes is a member of the editorial board of *Bioinformatics Advances*, ISCB - Oxford University Press.

Reviewer - reviewing activities All members of the team regularly review articles for international journals in Computational Biology or Bioinformatics.

11.1.4 Leadership within the scientific community

- Marie-Dominique Devignes is co-responsible of the Interoperability task force at the French Institute of Bioinformatics and participant of the ELIXIR Interoperability platform.
- Marie-Dominique Devignes proposed and co-organized a mini-symposium on Open Science and Interoperability at the *JOBIM* conference, 6-9 July 2021, Paris.
- Isaure Chauvot de Beauchêne co-organized a mini-symposium on "La modélisation structurale intégrative à l'ère des mégadonnées et de l'intelligence artificielle" at the *JOBIM* conference, 6-9 July 2021, Paris.
- Malika Smaïl-Tabbone was invited to chair a session at the *10th International Conference on Complex Networks and their Applications*, November 30 - December 2, 2021, Madrid.

11.1.5 Scientific expertise

Marie-Dominique Devignes reviewed applications grants for a call from Nouvelle Aquitaine Region.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Malika Smaïl-Tabbone is associate professor at the Université de Lorraine with a full service. She is co-responsible with Pascal Moyal of the IMSD track ("Ingénierie Mathématique pour la Science des Données") in the Applied Mathematics Master's degree at the Université de Lorraine. She is also member of the pedagogic team of the CMI BSE ("Cursus Master Ingénieur Biologie-Santé-Environnement") and also in charge of corporate relations.
- Sabeur Aridhi is associate professor at the Université de Lorraine with a full service. He is responsible for the major in IAMD ("Ingénierie et Applications des Masses de Données") at TELECOM Nancy.
- Marie-Dominique Devignes teaches every year 10 to 16h in the CMI BSE.
- Athenaïs Vaginay started an ATER position in september 2021.

11.2.2 Supervision

- Isaure Chauvot de Beauchêne supervised the 3A Mines internship of Alix Delannoy (co-supervised with Antoine Moniot and Yann Guermeur) and the M2 internship of Nathalie Bernard (co-supervised with Fariza Tahi and Eric Angel at IBISC, Evry). She also supervised the secondment of Anna Perez Rafols in the frame of the ITN RNAct.
- Marie-Dominique Devignes co-supervised with Isaure Chauvot de Beauchêne the internship of Karina Pats (exchange ITMO St-Petersbourg - Ecole des Mines Nancy) and with Anne Gégout-Petit (IECL) the M2 internship of Dalil Rouabah. She also supervised the MIASHS L3 internship of Valentin Retter and the CMI BSE L3 internship of Nathan Nourdin.
- Sabeur Aridhi supervised the M2 internship of Amal Stiti.

- Malika Smaïl-Tabbone co-supervised with Hélène Dumond the M2 internship of Nolwenn Lebourdais (affected to CRAN).
- Bernard Maigret supervised the M1 internship (initiation to research) of Nicolas Bombarde and Ugo Cottin.

11.2.3 Juries

- Marie-Dominique Devignes: reviewer in the PhD defense committee of Julie Lao, Université de Lorraine, 23 Feb 2021, examinator in the PhD defense committee of Louis Becquey, University of Evry, 6 Oct 2021, and reviewer in the PhD defense committee of Floris Chabrun, University of Angers, 7 Dec 2021.
- Marie-Dominique Devignes and Sabeur Aridhi: co-supervisors in the PhD defense committee of Bishnu Sarker, Université de Lorraine, 23 April 2021.
- Sabeur Aridhi: Co-supervisor in the PhD defense committee of Wissem Inoubli, University of Tunis, 7 Jan 2021.
- Malika Smaïl-Tabbone: invited member in the PhD defense committee of Nicolas Scalzitti, University of Strasbourg, 29 Sept 2021, co-supervisor in the PhD defense committee of Kevin Dalleau, Université de Lorraine, 23 Nov 2021.

11.3 Popularization

11.3.1 Internal or external Inria responsibilities

- Isaure Chauvot de Beauchêne was member of the hiring committee of the Inria competition for researcher recruitment.
- Isaure Chauvot de Beauchêne was member of two selection committees for associate professor positions at Université de Reims Champagne-Ardenne and Université de Lyon.

12 Scientific production

12.1 Major publications

[1] S. Z. Alborzi, A. Ahmed Nacer, H. Najjar, D. W. Ritchie and M. D. Devignes. 'PPIDomainMiner: Inferring domain-domain interactions from multiple sources of proteinprotein interactions'. In: *PLoS Computational Biology* 17.8 (2021), e1008844. DOI: 10.1371/journal.pcbi.1008844. URL: https://hal.archives-ouvertes.fr/hal-03435140.

[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] G. Preud'homme, K. Duarte, K. Dalleau, C. Lacomblez, E. Bresso, M. Smaïl-Tabbone, M. Couceiro, M.-D. Devignes, M. Kobayashi, O. Huttin, J. P. Ferreira, F. Zannad, P. Rossignol and N. Girerd. 'Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark'. In: *Scientific Reports* 11.1 (18th Feb. 2021), p. 4202. DOI: 10.1038/s41598-021-83340-8. URL: https://hal.univ-lorraine.fr/hal-03165272.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

12.2 Publications of the year

International journals

[18] S. Z. Alborzi, A. Ahmed Nacer, H. Najjar, D. Ritchie and M.-D. Devignes. 'PPIDomainMiner: Inferring domain-domain interactions from multiple sources of proteinprotein interactions'. In: *PLoS Computational Biology* 17.8 (2021), e1008844. DOI: 10.1371/journal.pcbi.1008844. URL: https://hal.archives-ouvertes.fr/hal-03435140.

[19] A. A. AWUSSI, E. Roux, C. Humeau, Z. HAFEEZ, B. Maigret, O. K. Chang, X. Lecomte, G. Humbert, L. Miclo, M. Genay, C. Perrin and A. Dary-Mourot. 'Role of the Sortase A in the Release of Cell-Wall Proteinase PrtS in the Growth Medium of Streptococcus thermophilus 4F44'. In: *Microorganisms* 9.11 (Nov. 2021), p. 2380. DOI: 10.3390/microorganisms9112380. URL: https://hal.archives-ouvertes.fr/hal-03537548.

[20] E. Bresso, P. Monnin, C. Bousquet, F.-E. Calvier, N.-C. Ndiaye, N. Petitpain, M. Smaïl-Tabbone and A. Coulet. 'Investigating ADR mechanisms with Explainable AI: a feasibility study with knowledge graph mining'. In: *BMC Medical Informatics and Decision Making* 21.1 (26th May 2021), p. 171. DOI: 10.1186/s12911-021-01518-6. URL: https://hal.inria.fr/hal-03240476.

[21] A. Hirtz, N. Lebourdais, F. Rech, Y. Bailly, A. Vaginay, M. Smaïl-Tabbone, H. Dubois-Pot-Schneider and H. Dumond. 'GPER agonist G-1 disrupts tubulin dynamics and potentiates temozolomide to impair glioblastoma cell proliferation'. In: *Cells* 10 (7th Dec. 2021), p. 3438. DOI: 10.3390/cells10123438. URL: https://hal.archives-ouvertes.fr/hal-03472082.

[22] M. K. Islam, S. Aridhi and M. Smaïl-Tabbone. 'An Experimental Evaluation of Similarity-Based and Embedding-Based Link Prediction Methods on Graphs'. In: *International Journal of Data Mining & Knowledge Management Process* 11 (30th Sept. 2021), pp. 1–18. DOI: 10.5121/ijdkp.2021.11501. URL: https://hal.inria.fr/hal-03540515.

[23] M. Kobayashi, O. Huttin, M. Magnusson, J. P. Ferreira, E. Bozec, A.-C. Huby, G. Preud'homme, K. Duarte, Z. Lamiral, K. Dalleau, E. Bresso, M. Smaïl-Tabbone, M.-D. Devignes, P. M. Nilsson, M. Leosdottir, J.-M. Boivin, F. Zannad, P. Rossignol and N. Girerd. 'Machine Learning-Derived Echocardiographic Phenotypes Predict Heart Failure Incidence in Asymptomatic Individuals'. In: *JACC: Cardiovascular Imaging* S1936-878X.21 (Sept. 2021), pp. 00556–8. DOI: 10.1016/j.jcmg.2021.07.004. URL: https://hal.univ-lorraine.fr/hal-03357064.

[24] H. Mezni, M. Sellami, S. Aridhi and F. B. Charrada. 'Towards big services: a synergy between service computing and parallel programming'. In: *Computing* 103.11 (Nov. 2021), pp. 2479–2519. DOI: 10.1007/s00607-021-00999-7. URL: https://hal.inria.fr/hal-03540412.

[25] G. Preud'homme, K. Duarte, K. Dalleau, C. Lacomblez, E. Bresso, M. Smaïl-Tabbone, M. Couceiro, M.-D. Devignes, M. Kobayashi, O. Huttin, J. P. Ferreira, F. Zannad, P. Rossignol and N. Girerd. 'Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark'. In: *Scientific Reports* 11.1 (18th Feb. 2021), p. 4202. DOI: 10.1038/s41598-021-83340-8. URL: https://hal.univ-lorraine.fr/hal-03165272.

[26] M. B. Veras, B. Sarker, S. Aridhi, J. P. Gomes, J. A. Macêdo, E. M. Nguifo, M.-D. Devignes and M. Smaïl-Tabbone. 'On the design of a similarity function for sparse binary data with application on protein function annotation'. In: *Knowledge-Based Systems* 238 (Feb. 2022), p. 107863. DOI: 10.1016/j.knosys.2021.107863. URL: https://hal.inria.fr/hal-03540409.

International peer-reviewed conferences

[27] A. Vaginay, T. Boukhobza and M. Smaïl-Tabbone. 'Automatic synthesis of boolean networks from biological knowledge and data'. In: *author's manuscript of OLA 2021, CCIS 1443 proceedings*. OLA 2021 - International Conference of Optimization and Learning. Catane, Italy, 21st June 2021. URL: https://hal.archives-ouvertes.fr/hal-03256693.

[28] A. Vaginay, T. Boukhobza and M. Smaïl-Tabbone. 'From quantitative SBML models to boolean networks'. In: *author's manuscript of Studies in Computational Intelligence 1016, SpringerComplex Networks & Their Applications XVolume 2, Proceedings of the TenthInternational Conference on ComplexNetworks and Their ApplicationsCOMPLEX NETWORKS 2021————-Rosa Maria Benito ·Chantal Cherifi · Hocine Cherifi ·Esteban Moro · Luis M. Rocha ·Marta Sales-Pardo Editors*. CNA 2021 - 10th International Conference on Complex Networks and their Applications. Madrid, Spain, 30th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03481396.

Conferences without proceedings

[29] A. Bazin, M. Couceiro, M.-D. Devignes and A. Napoli. 'A Hybrid Approach to Identifying the Most Predictive and Discriminant Features in Supervised Classification Problems'. In: ICCS 2021 - 26th International Conference on Conceptual Structures. Virtual, France, 15th Sept. 2021. DOI: 10.1007/978-3-030-86982-3_4. URL: https://hal.archives-ouvertes.fr/hal-03173406.

[30] A. Delannoy, A. Moniot, Y. Guermeur and I. Chauvot de Beauchêne. 'Feature extraction for the clustering of small 3D structures: application to RNA fragments'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris/Virtual, France, 6th July 2021, pp. 1–6. URL: https://hal.archives-ouvertes.fr/hal-03540927.

[31] M. K. Islam, S. Aridhi and M. Smaïl-Tabbone. 'Appraisal Study of Similarity-Based and Embedding-Based Link Prediction Methods on Graphs'. In: CDKP 2021 - 10th International Conference on Data Mining & Knowledge Management Process. London, United Kingdom: AIRCC Publishing Corporation, 24th July 2021, pp. 81–92. DOI: 10.5121/csit.2021.111106. URL: https://hal.archives-ouvertes.fr/hal-03540371.

[32] B. Sarker, M.-D. Devignes, G. Wolf and S. Aridhi. 'Prot-A-GAN: Automatic Protein Function Annotation using GAN-inspired Knowledge Graph Embedding'. In: ICML 2021 - Workshop on Computational Biology. Virtual, United States, 24th July 2021. URL: https://hal.inria.fr/hal-03541255.

[33] A. Vaginay, T. Boukhobza and M. Smaïl-Tabbone. 'Automatic synthesis of boolean networks from biological knowledge and data'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03481253.

[34] A. Vaginay, T. Boukhobza and M. Smaïl-Tabbone. 'From quantitative SBML to boolean networks'. In: CMSB 2021 - 19th conference on Computational Methods in Systems Biology, Bordeaux, France, 22nd Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03481267.

Scientific book chapters

[35] M. K. Islam, S. Aridhi and M. Smaïl-Tabbone. 'Simple negative sampling for link prediction in knowledge graphs'. In: *Complex Networks & Their Applications X*. Vol. 1016. Studies in Computational Intelligence. Springer International Publishing, 1st Jan. 2022, pp. 549–562. DOI: 10.1007/978-3-030-93413-2_46. URL: https://hal.archives-ouvertes.fr/hal-03540341.

Doctoral dissertations and habilitation theses

[36] W. Inoubli. 'Analysis and Mining of Large Dynamic Graphs: case of graph clustering'. Université de Tunis El Manar (Tunisie), 7th Jan. 2021. URL: https://hal.archives-ouvertes.fr/tel-03428615.

[37] B. Sarker. 'On Graph-Based Approaches for Protein Function Annotation and Knowledge Discovery'. Université de Lorraine, 23rd Apr. 2021. URL: https://hal.univ-lorraine.fr/tel-03274084.

Reports & preprints

[38] W. Inoubli, S. Aridhi, H. ·. Mezni, M. Maddouri and E. Mephu Nguifo. *A Distributed and Incremental Algorithm for Large-Scale Graph Clustering*. 2021. URL: https://hal.inria.fr/hal-02190913.

Other scientific publications

[39] S. Gulberti, C. Valencia-Schmitt, P. Villa, I. Chauvot de Beauchêne, 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'. In: Rencontres RARE 2021. Paris/virtuel, France, 14th Oct. 2021. URL: https://hal.univ-lorraine.fr/hal-03522390.

[40] A. Kravchenko, M. Smaïl-Tabbone, I. Chauvot de Beauchêne and S. J. De Vries. 'New strategy for optimizing knowledge-based docking parameters: application to ssRNA-protein docking'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03370998.

[41] A. Moniot, I. Chauvot de Beauchêne and Y. Guermeur. 'Agglomerative clustering of fragment 3D structures based on pairwise RMSD'. In: ISMB ECCB 2021. Virtual, France, 25th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03432682.

[42] A. Moniot, I. Chauvot de Beauchêne and Y. Guermeur. 'New clustering method to infer prototypes covering the 3D structures of nucleic acid fragments'. In: RECOMB 2021 - 25th International conference on research in computational molecular biology. Padova/Virtual, Italy, 29th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03432671.

12.3 Cited publications

[43] 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.

[44] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha and A. G. Murzin. 'SCOP2 prototype: a new approach to protein structure mining'. In: *Nucleic Acids Res* 42.Database issue (Jan. 2014), pp. D310–314. DOI: 10.1093/nar/gkt1242.

[45] A. Andreeva, E. Kulesha, J. Gough and A. G. Murzin. 'The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures'. In: *Nucleic Acids Res* 48.D1 (Jan. 2020), pp. D376–D382. DOI: 10.1093/nar/gkz1064.

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

[47] 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.

[48] M. Blum, H. Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar, L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman and R. D. Finn. 'The InterPro protein families and domains database: 20 years on'. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D344–D354. DOI: 10.1093/nar/gkaa977.

[49] 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.

[50] 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.

[51] 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.

[52] K. Dalleau. 'Une approche stochastique à base d'arbres aléatoires pour le calcul de dissimilarités : application au clustering pour diverses structures de données'. Theses. Université de Lorraine (Nancy), Nov. 2021. URL: https://tel.archives-ouvertes.fr/tel-03540627.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

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

[64] 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.

[65] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn and A. Bateman. 'Pfam: The protein families database in 2021'. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D412–D419. DOI: 10.1093/nar/gkaa913.

[66] 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.

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

[68] 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.

[69] D. N. Nicholson and C. S. Greene. 'Constructing knowledge graphs and their biomedical applications'. In: *Comput Struct Biotechnol J* 18 (2020), pp. 1414–1428. DOI: 10.1016/j.csbj.2020.05.017.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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, Sept. 2020, pp. 261–272. DOI: 10.1007/978-3-030-45385-5_24. URL: https://hal.inria.fr/hal-03025827.

[76] 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.

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

[78] 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.

[79] 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.

[80] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees and C. A. Orengo. 'CATH: increased structural coverage of functional space'. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D266–D273. DOI: 10.1093/nar/gkaa1079.

[81] 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.

[82] 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.

[83] 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.

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

[85] 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.