RESEARCH CENTRE

Inria Nancy - Grand Est Center

IN PARTNERSHIP WITH:

Université de Lorraine, CNRS

2022

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

Inria

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, Maternal leave until January 28, 2022, followed by 60% parental part-time until October 2022, and 80% parental part-time since then.]
- Hamed Khakzad [Inria, Chair, from Dec 2022, Junior Professor Chair]
- Bernard Maigret [CNRS, Emeritus]

Faculty Members

- Sabeur Aridhi [UL, Associate Professor]
- Malika Smaïl-Tabbone [UL, Associate Professor, HDR]

Post-Doctoral Fellow

- Dominique Mias-Lucquin [UL, until Sep 2022]

PhD Students

- Diego Amaya-Ramirez [Inria, until Sep 2022, UL since October 2022]
- Hrishikesh Dhondge [CNRS]
- Kamrul Islam [UL]
- Anna Kravchenko [CNRS]
- Athenaïs Vaginay [UL]

Technical Staff

- Antoine Moniot [UL, Engineer, from Oct 2022]

Administrative Assistants

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

Visiting Scientist

- Yasaman Karami [INSTITUT PASTEUR, from Nov 2022, Visiting scientist before joining the team in January 2023 as CRCN Inria.]

External Collaborator

- Taha Boukhobza [UL, Membre associé]

2 Overall objectives

2.1 Computational Challenges in Structural Biology

NB: This section has been remodeled since the death of Dave Ritchie in 2019.

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 [51, 63].

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

In the first axis, our main objective is to design, implement and test new KDD ("Knowledge Discovery in Databases") approaches to exploit specifically the structural information contained and sometimes hidden in many biological databases. These approaches will be oriented towards understanding molecular interactions in living organisms under physiological or pathological conditions.

In the second axis, our main objective is to propose new and fast methods to model the 3D structure of multi-component systems and characterize their dynamic behaviour. The challenge here is to integrate molecular flexibility into 3D models, thanks to molecular dynamics simulation and/or combinatorial approaches.

Finally, the complementarity of the two axes will be expressed through a common objective oriented towards the proposal of possible new treatments against diseases, based on the knowledge extracted and on the advances in 3D modeling of flexible molecular interactions. This objective will benefit from our network of biologist and clinician partners.

3 Research program

This section presents the current CAPSID research program. Several subjects initially present at the creation time (2015) or at last evaluation (2017) are no longer presented due to the death of Dave Ritchie.

3.1 Knowledge Discovery in Structural Databases

3.1.1 Context

In this axis, the CAPSID team develops methods related to knowledge discovery from databases (KDD, [37]). The diversity of biological databases and resources is such today that it is more and more difficult to consider each database independently from the others [54]. A limited subset of these resources is devoted to the 3D structure of biological objects (proteins, nucleic acids, glycanes...). Structural information is also contained in databases classifying protein domains as building blocks of proteins that can be reused in different proteins sharing the same function (Pfam, CATH and InterPro are well-known examples of such databases) [49, 59, 30]. There are millions of proteins across all living species but only tens of thousands of domains that are combined in proteins. Thus, complex tasks such as predicting protein function or interactions can be simplified when envisaged at the domain level.
Due to the great diversity of databases, Knowledge Graphs (KGs) are more and more used to represent and integrate biological information. There is no single definition of KGs as these graphs cover a large variety of domains and data representation contexts (for instance the GAFAM companies advertize various KG uses). The main feature that differentiates KGs from classical graphs is the fact that both nodes (or entities) and edges (or relations) in the graph are heterogeneous and belong to various types described in the KG schema (metagraph). The field of biological knowledge discovery in KG is expanding rapidly [47]. Most biological KGs today are developed for drug repurposing tasks (e.g. HetioNet[43] or DRKG). Clinicians are also very interested in network science carried out on rich knowledge graphs as a mean to interpret biomarker studies. However, there is still a need for curated, reliable biological KGs and for efficient knowledge discovery methods in KGs.

3.1.2 Knowledge discovery from protein structural databases

Concerning protein structural databases, we aim to explore novel classification paradigms exploiting existing resources about protein folds and domains [26, 27, 49, 59]. In particular it will be interesting to use Kpax, our structural alignment tool [55], to define domain-domain similarity matrices. A non-trivial issue with 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 KDD process leading from data to knowledge.

For example, protein domain classification is relevant for studying domain-domain interactions (DDI). Our previous work on Knowledge-Based Docking (KBDOCK, [39, 41]) 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.

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.

3.1.4 Knowledge discovery algorithms in large biological knowledge graphs

KGs are particularly useful and appropriate in biology, to represent and integrate the complex contents of biological databases[53]. We intend to design algorithms for leveraging information embedded in biological KGs (also known as complex networks). In biology, KGs 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.

3.2 Integrative Multi-Component Assembly and Modelling

3.2.1 Context

This axis deals with 3D protein structure and interactions. In fact, the long-lasting problem of predicting a 3D structure from a protein sequence has been solved in 2021 by the AlphaFold2 (DeepMind) [45] or RosettaFold methods [29]. This success, revealed in the CASP14 (Critical Assessment of Stucture Prediction) challenge, was possible not only thanks to AI methods but also because the amount of experimental 3D structures has reached a sufficient size in the Protein Data Bank (PDB). For the same type of reasons, the rigid docking problem (in which the bodies to dock are rigid) seems to be on the way to being solved as well [31, 36]. However, research is still required to address the problem of docking disordered proteins or flexible nucleic acids that will fold as they bind to proteins. This is the direction taken by the team since the arrival of Isaure Chauvot de Beauchêne, the inventor of a fragment-based approach for RNA docking onto proteins.
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 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 [35, 48, 50, 52]. In our experience, docking ensembles of NMA conformations do not give much improvement over basic FFT-based soft docking [62], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [40].

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 [28]. Typically, a CG force-field representation replaces the 5-15 atoms in each amino acid with 2-4 “pseudo-atoms” (each pseudo-atom represents few atoms of an amino-acid as a single bead). It then

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 [58]. Furthermore, this type of CG model effectively integrates many internal DOFs to build a smoother but still physically realistic energy surface [44]. We are currently developing a CG scoring function for RNA-protein docking by fragments assembly.

3.2.3 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 [34], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space, as initiated with the EROS-DOCK software [6, 57].

3.2.4 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 [32, 33].

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 want to develop new algorithms for the clustering of 3D conformations based on internal coordinates and on epsilon-net theory, in order to optimise the representativity and computational cost of the library.

In the future, we will improve the combinatorial algorithm used for reassembling the docked fragments using both experimental constraints and knowledge-based constraints pertaining from the research carried out in Axis 1.

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 [61]. 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

[25]. 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 [38, 56]. 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 [60]. 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, Bernard Maigret.

A third application domain has emerged in Axis 2 through the Inria-Inserm PhD thesis project 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 project 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

In structural bioinformatics and deep learning approaches, the computational costs are usually very high. The CAPSID team pays attention to use shared equipment (Platform MBI-DS4H, Grid5K) for running HPC (High Performance Computing) jobs as efficiently as possible. In particular, we use the "best effort" mode for distributed jobs.
Also, the CAPSID team is engaged in a collaboration with EMBRAPA (Brazilian Institute for Research in Agronomy, Brasilia) aimed at identifying environmentally friendly plant protection solutions, using virtual screening (see section 8.2.3).

6 Highlights of the year

6.1 Awards or success

- The team published an article in Journal of Biomedical Informatics (IF 6.3) about the lessons learnt from our participation to the RHU FIGHT-HF: Inductive database to support iterative data mining: Application to biomarker analysis on patient data in the Fight-HF project. This article is a good reflection of our advanced skills in both KDD approaches and molecular interaction data analysis, in the context of understanding diseases.

7 New software and platforms

7.1 New software

7.1.1 InteR3Mdb

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

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

Scientific Description: InteR3Mdb is a comprehensive database gathering and integrating public data on the 3D structures and sequences and on in vitro experiments, related to all instances known so far of a very conserved protein domain, the RNA Recognition Motif (RRM). Special effort has been made to enter in the database all available and curated information about the interactions of these RRM instances with RNA.

Functional Description: InteR3Mdb is a comprehensive database built on a relational data model. As a tool, it comprises a web-interface offering filtering and multicriteria search functionalities about RRMs, as well as an API interface.

Release Contributions: InteR3Mdb V1.0 is released with an extended public user interface offering filtering and multi-criteria querying functionalities. The user manual, data dictionary and help documentation are now fully available as well as the main database content statistics (https://inter3mdb.loria.fr)

News of the Year: During year 2022, the InteR3Mdb data model has evolved to integrate results about binding affinities between RRM-containing proteins and RNA. The list of represented RRM domains has been updated. A public interface (https://inter3mdb.loria.fr) has been created with filtering and multicriteria search functionalities. Documentation (Data Dictionary, Statistics) is available on the web site.

URL: `https://inter3mdb.loria.fr`

Contact: Hrishikesh Dhondge

7.1.2 ProtNAff

Name: Protein - Nucleic Acids Filters and Fragments

Keywords: Structural alphabet, Structural Biology, Nucleic Acids

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

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

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

Publication: hal-02393039

Contact: Isaure Chauvot de Beauchêne

Participants: Isaure Chauvot de Beauchêne, Antoine Moniot, Sjoerd De Vries

7.1.3 DISCAN-2022

Name: Distributed and Incremental Structural Clustering Algorithm for Networks

Keywords: Big data, Clustering, Unsupervised graph clustering SCDG, Distributed computing

Functional Description: This is a distributed implementation of a Distributed and Incremental Structural Clustering Algorithm for Networks called DISCAN. For more details see the paper "A distributed and incremental algorithm for large-scale graph clustering" by Wissem Inoubli, Sabeur Aridhi, Haithem Mezni, Mondher Maddouri, Engelbert Mephu Nguifo. ⟨10.1016/j.future.2022.04.013⟩.

URL: https://github.com/inoubliwissem/remote-master

Publication: hal-03659549

Contact: Wissem Inoubli

Partner: Université Clermont Auvergne

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 at the origin of the creation of the LORIA MBI-DS4H research platform that provides a shared environment to the CAPSID and ORPAILLEUR teams for running distributed intensive computation. This platform is also the place for optimizing codes that can be run later on Grid 5K or on the Jean-Zay supercalculator. Moreover, the platform offers opportunities for newcomers in the team to get trained to good practices in development and in sharing code and data.

The technical support of the patform is ensured by the LORIA SISR (Service d'Ingénierie en Soutien de la Recherche) via a private project on gitlab.

8 New results

8.1 Axis 1 : Knowledge Discovery in Structural Databases

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

8.1.1 Biomedical Knowledge Discovery

In the context of our collaboration with clinicians at the CHRU Nancy in the RHU FIGHT-HF project (2015-2021 ; see preceding report), we have proposed an approach based on an inductive database to support iterative data mining [8]. In this approach, we extend the KDD process model with the help of an inductive database and we design the first generic model of Inductive Clinical DataBase (ICDB) aimed at hosting both patient data and learned models. We report experiments conducted on patient data within the FIGHT-HF program. The ICDB approach allows us to identify biomarker combinations, which are specific and predictive of heart fibrosis phenotype and provide hypotheses about the underlying mechanisms. Two main scenarios were considered, a local-to-global KDD scenario and a trans-cohort alignment scenario. This promising proof of concept enables us to draw the contours of a next-generation Knowledge Discovery Environment (KDE).
In parallel, we continue to exploit the network science platform developed during the FIGHT-HF project. Queries on the Heart Failure Graph Knowledge Box (HF-GKBox) have provided useful information to clinicians for interpreting their results concerning the association of two biomarkers (NT-proBNP and Stem Cell Factor) present in blood plasma with cardiovascular outcomes in end-stage hemodialysis patients with renal disease [15].

In the context of the PraktikPharma ANR project coordinated by Adrien Coulet (HeKA team), we investigated adverse drug reaction mechanisms with knowledge graph mining [23].

8.1.2 Graph-based Approaches for Machine Learning and Protein Annotation

The developments conducted by Bishnu Sarker to improve his method for protein function annotation (GrAPFI for Graph-based Automatic Protein Function annotation) have been published. GrAPFI is based on label propagation in graphs. It relies on a graph whose nodes represent proteins and whose edges are weighted by the domain similarity between proteins. 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. It was shown that the post-processing step also improves other methods of protein function annotation [16].
Similarity functions are essential for label propagation algorithms in graphs. As an extension of our Tempo-Graphs project (2019-2021) with our Brazilian partners at Federal University of Ceara (UFC, Fortaleza), we developed a similarity function for sparse binary data with application on protein function annotation [19]. 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.
Moreover, through the past co-supervision by Sabeur Aridhi of Wissem Inoubli's PhD at the University of Tunis (2018-2021), the team has acquired recognized expertise in big data methods for large graphs and contributed to the development of a distributed, incremental algorithm for large-scale graph clustering [10]. The corresponding software DISCAN-2022 is described above (see section New Software).

8.1.3 Knowledge graph mining with embedding-based methods

In the context of Md Kamrul Islam's PhD project, we addressed the problem of link prediction in large knowledge graphs (KGs) using KG embedding methods. These methods aim to learn low dimensional vector representations of entities and relations in a KG. Such representations (in a latent space) facilitate

the link prediction task, which serves inference and completion in a KG. In this context, it is important to achieve both an efficient KG embedding and explainable predictions. During learning of efficient embeddings, sampling negative triples is an important step as KGs only come with observed positive triples. We proposed an efficient 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 [11].
As for explainability, it actually constitutes a thriving research question especially when it comes to analyse KGs with their rich semantics rooted in description logics. Hence, we proposed a new rule mining method which exploits the learned embeddings [11].
We evaluated our SNS sampling method plugged to several KG embedding models by calculating the performance of the link prediction task on well-known datasets. Experimental results show that the SNS improves the prediction performance of KG embedding models, and outperforms the existing sampling methods. To assess the performance of our rule mining method with and without SNS, we mine and evaluate rules on three popular datasets. The extracted rules are evaluated as knowledge nuggets extracted from the KG and also as a support for explainable link prediction. The achieved results are good and open the way to many improvements and new perspectives. Md Kamrul Islam has defended his PhD on December 16, 2022.

8.1.4 Biological network modeling

Boolean Networks (BNs) refer to 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. The BioModels database is a well-known repository of peer-reviewed models represented in the Systems Biology Markup Language (SBML). Most of these models are quantitative, but in some use cases, qualitative models—such as BNs—are better suited. In the context of Athénais Vaginay's PhD project, we proposed SBML2BN, a pipeline dedicated to the automatic transformation of quantitative SBML models to Boolean networks [18]. Our approach takes advantage of several SBML elements (reactions, rules, events) as well as a numerical simulation of the concentration of the species over time to constrain both the structure and the dynamics of the Boolean networks to synthesise. Finding all the BNs complying with given structure and dynamics was formalised as an optimisation problem formulated in the answer-set programming framework.
We ran SBML2BN on more than 200 quantitative SBML models, and we could construct Boolean networks which are compatible with the structure and the dynamics of the SBML models [18]. A more recent work relies on abstract simulation of a chemical reaction network (CRN) to avoid the tricky binarization task [21]. Athénais Vaginay will defend her PhD in May 2023.

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, Diego Amaya Ramirez.

8.2.1 Inferring epsilon-nets of RNA 3D-fragments

Our fragment-based approach to dock ssRNA on proteins requires libraries of 3D conformations of RNA fragments. Last year, we reported ProtNAff, a python-based software for (i) the automated parsing and annotation of protein-RNA experimental structures, (ii) the selection of (parts of) such structures by customized queries, (iii) the creation of fragment libraries. Since then, we have applied it to characterize the conformational specificity of RNA in different states: single-stranded or double-stranded, unbound or bound to proteins, of different sequences, etc. The ProtNAff software and these analyses were published in a peer-reviewed journal [14].

The RNA fragment libraries must approximate all possible local RNA conformations within a given precision, but be of small enough cardinality to be usable for combinatorial assemblies of fragments. Such a set of well chosen prototypes can be obtained by clustering all fragments of experimentally solved structure, using a suitable clustering algorithm and a measure of dissimilarity between fragments. When the representativeness criterion is a distance, the problem reduces to that of inferring an epsilon-net [1] (ε-net) of minimal cardinality for a finite set of points in a metric space. After the construction of the ε-net, each point P_i from the initial set should be at a distance $d_i < \varepsilon$ from at least one point in the ε-net. The minimal cardinality that can be achieved is called the ε-covering number.
In collaboration with Yann Guermeur from team ABC (LORIA), we investigated methods to construct minimal ε-nets from a given set of points and a given distance measure. Although the literature provides us with upper bounds on the covering numbers [2], the proofs are non constructive. To the best of our knowledge, there was no (efficient) algorithm to compute ε-nets of small cardinality in the framework of interest. We have therefore developed an algorithm to derive ε-nets, that operates in a reproducing kernel Hilbert space [3], by combining two well-known tools of empirical inference: the hierarchical agglomerative clustering and the computation of minimum enclosing balls. Tested on a well-known set of images used for benchmarking classification tools, it produces ε-nets whose cardinality is smaller than those obtained with state-of-the-art methods. This work has been presented at the WSOM international conference [20].

We then applied this approach on RNA fragments, with some arrangements. A commonly used measure of dissimilarity in structural biology is the root mean squared deviation (RMSD) of atomic positions, whose exact computation requires a pairwise structural superposition of the 3D structures. But this superposition is highly time-consuming and not applicable for a very large initial set of fragments (typically 10^4 to 10^5 in our cases). In our approach, we first approximate the real RMSD by the RMSD after superposition on one random fragment of the set, which provides an upper bound of the real RMSD. When a ball calculation is to be performed, all the fragments involved are superimposed again on a common reference structure. To know which balls are really useful to compute, bounds are used (to limit superpositions). This allowed us to reduce by a factor 4 the size of our fragment libraries, compared to the results obtained by the star-shape clustering implemented in the first version of ProtNAff. This work was carried out with ABC in the context of Antoine Moniot's PhD, co-supervised by Yann Guermeur and Isaure Chauvot de Beauchêne, and successfully defended on December 12, 2022.

8.2.2 Modeling and design of RNA-RRM complexes

RNA docking with stacking constraints. Our H2020 ITN project RNAct aims at designing new RNA-binding proteins based on the evolutionary conserved protein domain[4], called RNA Recognition Motif (RRM). In this context, we have created in 2020 the InteR3Mdb database that contains all known 3D structures of RRMs and RRM-RNA complexes. It allowed us to infer, for each amino acid at every evolutionary conserved position in the sequence[5], the propensity to bind a given nucleotide type (A/C/U/G) with a given interaction type (electrostatic, Hydrogen-bond, etc). We are now using this resource to derive constraints for the modeling of an RNA bound to a given RRM. As a first step, we concentrated on the $\pi-\pi$ stacking interactions occurring between a base of the RNA and the aromatic ring of an amino-acid at evolutionary conserved positions in an RRM. We inferred, from superposition and clustering of all existing RNA-RRM structures, the main possible positions of a stacking RNA base toward an RRM structure. Then, on a test-case RRM structure, we modeled the RNA binding positions by docking (i.e. sampling low-energy positions for RNA fragments) with an energy penalty added to the RNA-protein force-field for deviations from the reference position. The approach was tested on 12 experimental RRM-RNA

[1] An epsilon-net in computational geometry is the approximation of a general set by a collection of simpler subsets. In probability theory it is the approximation of one probability distribution by another.
[2] The so-called extended Sauer-Shelah lemmas [42].
[3] A reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions in which point evaluation is a continuous linear functional.
[4] An evolutionary conserved protein domain is a protein domain for which the aligned sequences do not differ much across all possible living species.
[5] An evolutionary conserved position in a protein sequence refers to the position in a sequence alignment of all known instances of this protein across all possible living organisms, where the same amino acid is always or nearly always found.

structures: it retrieved correct positions (RMSD < 1 Å from the target position) for all test cases, versus only 1 test case by docking without constraints.

Cross-mapping of protein domain structural instances. To address the issue of inconsistencies of protein classification in different databases, we developed a tool named Cross-Mapper for domain Structural instances (CroMaSt) that is a workflow to identify and assign a confidence level to each experimental 3D structure of a given domain. This workflow was primarily developed for RRMs, but can easily be adapted to any structural domain. The workflow has been formalized using the Common Workflow Language (CWL) [6]. Its rationale is based on the cross-mapping of PDB (Protein Data Bank) entries retrieved as RRM, in the Pfam and CATH domain classifications. Entries that are mapped in both classifications are considered as the "core" RRMs. Those that are specific to only one classification are further analyzed using structural alignment, in order to determine whether they are RRM-like or false RRMs [24]. The workflow can be used to create datasets for specific domains with a good confidence level, which are useful for characterizing domain structural diversity or for further analyses such as machine learning, evolutionary studies or synthetic biology. An article on CroMaSt and the rich diversity of RRM domains has been submitted to Bioinformatics in November 2022 (under review).

Empirical inference of an RNA-protein energy function. In collaboration with Sjoerd de Vries [7], 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, with 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. They were initially tailored for double-stranded RNA, and their performance is poor on single-stranded RNA (ssRNA). One main goal of Anna Kravchenko's PhD is to optimize those parameters for the discrimination of correct models for a given ssRNA-protein complex. To achieve that, we have set up a novel "histogram-based" approach. For each pair of bead types, we build a log-odds histogram of the occurrences of distances (discretized into bins) in correct/incorrect models from docking runs on a training set. We then use these histograms, which correspond to the residual error of the energy function, to score each individual model in docking runs on a test set. We tested this corrected approach on a benchmark of 131 complexes of known structures of ssRNA trinucleotides bound to 56 proteins and Anna Kravchenko presented the results as an oral communication at the 35th Rhine-knee Regiomeeting for Structural Biology (5-7 October 2022, Gerardmer, France). In perspective, if we can find a way to transform a set of histograms into a set of docking parameters then we can use this approach to also improve the sampling step, by increasing the raw number of correct models among all generated models.

8.2.3 3D Modeling of protein complexes - Virtual Screening

Modeling DNA-protein complexes to fight antibiotic resistance. In the context of the CITRAM project (collaboration with the DynAMic and LPCT labs), we study a new type of relaxase (RelSt3) as a key protein for horizontal DNA transfer, one of the processes responsible for the spread of antibiotic resistance in bacteria. At the very beginning of this process, DNA interacts with the relaxase both as a classical double-helix and as a transient single-stranded state. We therefore developped a ssDNA docking tool to model the interaction of the relaxase RelSt3 with ssDNA. To benchmark this tool, we created the first dataset of ssDNA-protein structures. From the public general database of experimental structures (PDB), we extracted, processed and clustered 284 complexes containing protein-bound ssDNA. A main difficulty in docking is the conformational differences between the unbound and the DNA-bound protein structures, the latter being not predictable from the former. Using many unbound structures for docking can increase the chance that at least one is close enough to the bound one to provide accurate models. Therefore, we also retrieved all unbound structures from the Protein Data Bank for each protein of our 284

[6] https:www.commonwl.org

[7] INSERM engineer who joined the SISR (Service d'Ingénierie et de Soutien à la Recherche) department at the LORIA as a CNRS engineer in July 2022

ssDNA-protein structures. We then used this dataset (i) to compare intrinsic and extrinsic conformational variability, i.e. structural changes induced or not by ssDNA binding, and (ii) to prove the advantage of using multiple protein structures in ssDNA docking [13].
In parallel, we modeled the interaction of the HtH domain[8] of RelSt3 with double-helix DNA, using known structures of homologous complexes and molecular dynamics simulations. Our models identified key residues for interaction with DNA, which were confirmed by the loss of relaxase activity in vitro after mutating those residues. Part of this work has been published in the Nucleic Acids Research journal, with our microbiologist partners [12].
COVID-19 spike binding to ACE2. In collaboration with Laurent Vuillon and Aria Gheeraert at the LAMA (Laboratoire de Mathématique, CNRS-Université Savoie Mont Blanc) and with Laurent Chaloin and Olivier Moncorgé at the IRIM (Institut de Recherche en Infectiologie de Montpellier, CNRS-Université de Montpellier), we investigated disparities between the SARS-CoV-2 wild-type and five variants that emerged at the end of 2020, focusing on the Spike/ACE2 complex and using experimental structures and molecular dynamics simulations. Dihedral angle PCA and dynamical perturbation networks showed specificities in the Spike dynamics and in the Spike/ACE2 interface respectively, in 4 and 3 variants compared to wild-type respectively. Atomic contacts PCA shows how the L452R and T478K mutations act synergistically on neighboring residues to provoke drastic changes compared to the wild-type behavior [9].
Structural basis of donor-specific antibody response in graft rejection. In the context of the Inria-Inserm PhD project of Diego Amaya Ramirez, in collaboration with the Immunology and Histocompatibility Laboratory at the APHP Saint-Louis Hospital in Paris, we study the structural differences between donor and recipient HLA proteins to understand and possibly predict the immune response triggered in the recipient, which can result in rejection of the transplanted organ. A dataset of 207 HLA 3D structures has been created, both from PDB and AlphaFold predictions. Molecular dynamics runs (10 ns) have been run to simulate protein surface flexibility. These runs are analyzed at the level of single residues or surface patches centered on such residues. Single residues are selected as "confirmed eplets" (polymorphic amino-acids responsible of generating donor-specific antibodies[9]) or "non eplets" (conserved amino-acids in all HLA proteins, assumed not to induce any donor-specific antibody response). Based on a recent publication suggesting that antibodies tend to bind to less flexible regions on the surface of the proteins [46], we studied the flexibility of HLA surface along molecular dynamics runs, at eplet versus non eplet positions. Our results reveal a significant difference which could indicate that HLA-specific antibodies would also preferentially bind to less flexible regions [22]. The datasets built during this preliminary study constitute the basis for further machine learning approaches helping to discriminate between compatible and non compatible recipient-donor HLA pairs.
Virtual screening Virtual screening of small molecules is an essential part of the team's expertise that is often called upon by external partners in biology. In collaboration with Brazilian partners from EMBRAPA, we developped a computational strategy to identify new efficient and environmentally-friendly plant protection solutions. Anti-mycotoxin compounds were isolated by virtual screening using as a target a key enzyme specific of the pathogen fungus [7]. In parallel, we participated in an innovative study with colleagues at the University of Maringa in which deep learning methods were combined with classical ligand- and target-based chemoinformatic methods to identify new drugs against the tuberculosis pathogen [17].

9 Bilateral contracts and grants with industry

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

[8] HtH : Helix turn Helix. Protein structural motif composed of two α helices joined by a short strand of amino acids, that binds to the major groove of DNA
[9] The EpRegistry database currently registers all known HLA eplets, confirmed or predicted

10 Partnerships and cooperations

10.1 International initiatives

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

FlexMol

Title: Algorithms for Multiscale Macromolecular Flexibility

Duration: 2019 - 2022

Coordinator: Pablo Chacón (pablo@chaconlab.org)

Partners:

- Sergei Grudinin, Laboratoire Jean Kuntzmann, Grenoble (France)
- Pablo Chacón, Rocasolano Institute of Physical Chemistry (IQFR-CSIC), Madrid, Spain (Espagne)

Inria contact: Marie-Dominique Devignes

Summary: The goal of this collaboration was to mutually explore novel computational techniques for emerging problems in structural biology and bioinformatics related to molecular flexibility. Molecular flexibility is essential to link structure and function of many biological macromolecules. Changes in protein conformation play a vital role in biochemical processes, from biopolymer synthesis to membrane transport. Many proteins can drastically alter their architecture and display considerable interdomain flexibility, as found in their 3D structures. For example, proteins rely on flexibility to respond to environmental changes, ligand binding and chemical modifications. Also, protein flexibility is tightly bound to their stability and is fundamental for drugs to exert biological effects.
Thus, one of the main challenges in the field of computational structural biology is to predict and explain molecular flexibility and corresponding conformational changes. For example, currently there are no methods that can reliably predict structural changes in proteins upon their binding. However, these are crucial to predict the structure of protein complexes with large conformational changes upon binding. To give another example, flexibility of the protein binding pocket is the major hurdle in reliable prediction of protein-ligand interactions for computer-aided drug design. Finally, intrinsic flexibility of macromolecules is nowadays the limiting factor for high-resolution experimental structure determination.

10.2 European initiatives

10.2.1 H2020 projects

ITN RNAct

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

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

Duration: October 2018 - October 2022

Coordinator: Wim Vranken (Vrije University Brussels, Belgium)

Inria contact: Isaure Chauvot de Beauchêne

Partners:

- Vrije University Brussels (Belgium)

- 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)
- Dynamics Biosensors Gmbh (Germany)

Summary: The RNAct International Training Network (ITN) 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

ANR EPIHLA

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

Title: HLA compatibility in organ transplantation : from antigens to epitopes (EPIHLA)

Duration: October 2022-October 2025

Coordinator: Pr. Jean-Luc Taupin (Inserm U976, Saint-Louis Hospital, Paris)

Inria contact: Marie-Dominique Devignes

Partner Institutions:
- Inserm U976 IRSL Saint-Louis Hospital (Paris)
- LORIA CNRS (Nancy)
- INSERM U1016 Cochin Institute (Paris)
- CNRS U144 Institut Curie (Paris)

Summary: The EPIHLA project has two major aims. (1) It aims at correctly representing HLA molecule 3D structure and superimposing predicted conformations in order to identify 3D differences that could constitute epitopes and eplets, targets of donor-specific antibodies. (2) It aims at developing the capacity to isolate and clone anti-HLA antibody genes from patients' B lymphocytes. The results will provide decisive new information on the understanding of humoral alloreactivity and will make it possible to better anticipate transplant rejection. This project is partially based on the Inria-Inserm PhD project of Diego Amaya Ramirez (2019-2022). This PhD ("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 Pr. Jean-Luc Taupin.

11 Dissemination

Participants: Marie-Dominique Devignes, Malika Smaïl-Tabbone, Sabeur Aridhi, Isaure Chauvot de Beauchêne, Athénaïs Vaginay, Diego Amaya Ramirez, Anna Kravchenko, Hrishikesh Dhondge.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- Isaure Chauvot de Beauchêne was a member of the organizing committee of the Bioinfo-3D Mini-Symposium at JOBIM 2022.

11.1.2 Scientific events: selection

Member of conference program committees

- Marie-Dominique Devignes : ACM-BCB 2022, IWBBIO 2022.
- Isaure Chauvot de Beauchêne : JOBIM 2022, RECOMB2023 (subreviewing).
- Malika Smaïl-Tabbone : IDA 2023.

11.1.3 Journals

Member of editorial boards

- Marie-Dominique Devignes is an invited member of the editorial board of Bioinformatics Advances, Oxford Press.

Reviewing activities

- Members of the team have reviewed papers for Frontiers in Molecular Biosciences, Computational Structural Biology Journal, MethodsX.

11.1.4 Leadership within the scientific community

- Marie-Dominique Devignes is coordinating the Interoperability working group at the Institut Français de Bioinformatique with a bioinformatician colleague from Nantes (Alban Gaignard, Institut du Thorax, Plateforme BiRD). This working group is a member of the ELIXIR Interoperability platform. She represents IFB for this part of its activity in meetings at the national and international levels.

11.1.5 Scientific expertise

- Marie-Dominique Devignes reviewed grant applications for the European Science Foundation (call FWO-Odysseus 2022) and for BPI-France (call i-demo 2022).
- Malika Smaïl-Tabbone is enrolled as an expert in data science and AI in two European projects coordinated by a leading clinician (Pr. Magnus Bäck): an HORIZON Europe project entitled CARE-IN-HEALTH (CARdiovascular Resolution of Inflammation to promote HEALTH) and an ERA PerMed project entitled OmegaPerMed (Optimizing omega-3 supplementation to resolve inflammation in a personalized medicine cardiovascular disease prevention). She is responsible for the design and execution of the data science- and AI-related tasks in both projects.
- Malika Smaïl-Tabbone and Marie-Dominique Devignes contributed to the elaboration of the interdisciplinary LUE program TRAVEL (Co-morbidiTies, tRAjectories of liVEs and Longevity), which gathers the three digital science laboratories in Nancy together with Nancy Hospital and research laboratories in public health, biology and social sciences.

11.1.6 Research administration

- Marie-Dominique Devignes is responsible for the Transverse Axis "Digital Health" at the LORIA.
- Marie-Dominique Devignes was member of the hiring committee for an assistant professor (Maître de Conférences) at Université Paris-Saclay in May 2022.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Malika Smaïl-Tabbone is an 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 a 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 an assistant 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 had an ATER position from September 2021 to August 2022.

- Diego Amaya Ramirez started an ATER position in October 2022.

11.2.2 Supervision

- PhD: Antoine Moniot. "Modélisation des complexes ARN/protéines par assemblage de fragments structuraux". Co-supervised by Isaure Chauvot de Beauchêne and Yann Guermeur (ABC team, Loria). PhD defended on December 12, 2022. Available soon at https:theses.hal.science.

- PhD: Md Kamrul Islam. "Explainable link prediction in large complex graphs - application to drug repurposing". Co-supervised by Malika Smaïl-Tabbone and Sabeur Aridhi. Thesis defended on December 16, 2022. Available soon at https:theses.hal.science.

- PhD in progress: Athenaïs Vaginay, October 2018, co-supervised by Malika Smaïl and Taha Boukhobza (CRAN).

- PhD in progress: Diego Amaya Ramirez, October 2019, co-supervised by Marie-Dominique Devignes and Jean-Luc Taupin (Inserm, Hôpital Saint-Louis, Paris).

- PhD in progress: Anna Kravchenko, October 2019,co-supervised by Isaure Chauvot de Beauchêne and Malika Smaïl-Tabbone.

- PhD in progress: Hrishikesh Dhondge, October 2019, co-supervised by Isaure Chauvot de Beauchêne and Marie-Dominique Devignes.

11.2.3 PhD thesis juries

- ALFERKH Lina, 10 February 2022, Université Paris-Saclay (Isaure Chauvot de Beauchêne, examiner)

- PETRELIS Alexandros, 29 April 2022, Université de Lorraine (Marie-Dominique Devignes, examiner)

- VERAS Marcelo, 24 June 2022, Federal University of Ceará, Brazil (Sabeur Aridhi, examiner)

- AKID-ZEKRI Hajer, 12 December 2022, co-tutelle Université de Strasbourg and Université de Sfax (Malika Smaïl-Tabbone, reviewer)

11.2.4 Other juries

- Malika Smaïl-Tabbone is member of the Capes NSI (Numérique et Sciences Informatiques) jury since 2022.

11.3 Popularization

11.3.1 Articles and contents

Two PhD students from the RNAct project have produced short presentation videos now available on YouTube (Anna Kravchenko and Hrishikesh Dhondge).

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] E. Bresso, J.-P. Ferreira, N. Girerd, M. Kobayashi, G. Preud'homme, P. Rossignol, F. Zannad, M.-D. Devignes and M. Smaïl-Tabbone. 'Inductive database to support iterative data mining: Application to biomarker analysis on patient data in the Fight-HF project'. In: *Journal of Biomedical Informatics* 135 (Nov. 2022), p. 104212. DOI: 10.1016/j.jbi.2022.104212. URL: https://hal.univ-lorraine.fr/hal-03805671.

[3] K. Islam, S. Aridhi and M. Smaïl-Tabbone. 'Negative Sampling and Rule Mining for Explainable Link Prediction in Knowledge Graphs'. In: *Knowledge-Based Systems* 250 (17th Aug. 2022), p. 109083. DOI: 10.1016/j.knosys.2022.109083. URL: https://hal.archives-ouvertes.fr/hal-03684205.

[4] A. Moniot, Y. Guermeur, S. J. de Vries and I. Chauvot de Beauchêne. 'ProtNAff: Protein-bound Nucleic Acid filters and fragment libraries'. In: *Bioinformatics* 38.162022-07-01 (2022), pp. 3911–3917. DOI: 10.1093/bioinformatics/btac430. URL: https://hal.science/hal-03765772.

[5] 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.

[6] 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.

12.2 Publications of the year

International journals

[7] V. Atanasova, E. Bresso, B. Maigret, N. F. Martins and F. Richard-Forget. 'Computational Strategy for Minimizing Mycotoxins in Cereal Crops: Assessment of the Biological Activity of Compounds Resulting from Virtual Screening'. In: *Molecules* 27.8 (Apr. 2022), p. 2582. DOI: 10.3390/molecules27082582. URL: https://hal.inria.fr/hal-03647997.

[8] E. Bresso, J.-P. Ferreira, N. Girerd, M. Kobayashi, G. Preud'homme, P. Rossignol, F. Zannad, M.-D. Devignes and M. Smaïl-Tabbone. 'Inductive database to support iterative data mining: Application to biomarker analysis on patient data in the Fight-HF project'. In: *Journal of Biomedical Informatics* 135 (Nov. 2022), p. 104212. DOI: 10.1016/j.jbi.2022.104212. URL: https://hal.univ-lorraine.fr/hal-03805671.

[9] A. Gheeraert, L. Vuillon, L. Chaloin, O. Moncorgé, T. Very, S. Perez, V. Leroux, I. Chauvot de Beauchêne, D. Mias-Lucquin, M.-D. Devignes, I. Rivalta and B. Maigret. 'Singular Interface Dynamics of the SARS-CoV-2 Delta Variant Explained with Contact Perturbation Analysis'. In: *Journal of Chemical Information and Modeling* 62.12 (27th June 2022), pp. 3107–3122. DOI: 10.1021/acs.jcim.2c00350. URL: https://hal.inria.fr/hal-03708020.

[10] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri and E. Mephu Nguifo. 'A distributed and incremental algorithm for large-scale graph clustering'. In: *Future Generation Computer Systems* 134 (Sept. 2022), pp. 334–347. DOI: 10.1016/j.future.2022.04.013. URL: https://hal.inria.fr/hal-03659549.

[11] K. Islam, S. Aridhi and M. Smaïl-Tabbone. 'Negative Sampling and Rule Mining for Explainable Link Prediction in Knowledge Graphs'. In: *Knowledge-Based Systems* 250 (17th Aug. 2022), p. 109083. DOI: 10.1016/j.knosys.2022.109083. URL: https://hal.archives-ouvertes.fr/hal-03684205.

[12] H. Laroussi, Y. Aoudache, E. Robert, V. Libante, L. Thiriet, D. Mias-Lucquin, B. Douzi, Y. Roussel, I. Chauvot de Beauch Êne, N. Soler and N. Leblond-Bourget. 'Exploration of DNA processing features unravels novel properties of ICE conjugation in Gram-positive bacteria'. In: *Nucleic Acids Research* 50.14 (12th Aug. 2022), pp. 8127–8142. DOI: 10.1093/nar/gkac607. URL: https://hal.archives-ouvertes.fr/hal-03881644.

[13] D. Mias-Lucquin and I. Chauvot de Beauchêne. 'Conformational variability in proteins bound to single-stranded DNA: a new benchmark for new docking perspectives'. In: *Proteins - Structure, Function and Bioinformatics* 90.3 (Mar. 2022), pp. 625–631. DOI: 10.1002/prot.26258. URL: https://hal.archives-ouvertes.fr/hal-03821504.

[14] A. Moniot, Y. Guermeur, S. J. de Vries and I. Chauvot de Beauchêne. 'ProtNAff: Protein-bound Nucleic Acid filters and fragment libraries'. In: *Bioinformatics* 38.162022-07-01 (2022), pp. 3911–3917. DOI: 10.1093/bioinformatics/btac430. URL: https://hal.archives-ouvertes.fr/hal-03765772.

[15] P. Rossignol, K. Duarte, E. Bresso, Å. A, M. D. Devignes, N. Eriksson, N. Girerd, R. Glerup, A. Jardine, H. Holdaas, Z. Lamiral, C. Leroy, Z. Massy, W. März, B. Krämer, P. Wu, R. Schmieder, I. Soveri, J. Christensen, M. Svensson, F. Zannad and B. Fellström. 'NT-proBNP and stem cell factor plasma concentrations are independently associated with cardiovascular outcomes in end-stage renal disease hemodialysis patients'. In: *European Heart Journal Open* 2.6 (1st Nov. 2022). DOI: 10.1093/ehjopen/oeac069. URL: https://hal.univ-lorraine.fr/hal-03925099.

[16] B. Sarker, N. Khare, M.-D. Devignes and S. Aridhi. 'Improving automatic GO annotation with semantic similarity'. In: *BMC Bioinformatics* 23.S2 (Dec. 2022), p. 433. DOI: 10.1186/s12859-022-04958-7. URL: https://hal.inria.fr/hal-03915205.

[17] J. V. P. D. Souza, E. S. Kioshima, L. S. Murase, D. d. S. Lima, F. A. V. Seixas, B. Maigret and R. F. Cardoso. 'Identification of new putative inhibitors of Mycobacterium tuberculosis 3-dehydroshikimate dehydratase from a combination of ligand- and structure-based and deep learning in silico approaches'. In: *Journal of Biomolecular Structure and Dynamics* (23rd Feb. 2022), pp. 1–10. DOI: 10.1080/07391102.2022.2042389. URL: https://hal.inria.fr/hal-03591628.

[18] A. Vaginay, T. Boukhobza and M. Smaïl-Tabbone. 'From quantitative SBML models to Boolean networks'. In: *Applied Network Science* 7.1 (Dec. 2022), p. 73. DOI: 10.1007/s41109-022-00505-8. URL: https://hal.archives-ouvertes.fr/hal-03902922.

[19] M. Veras, B. Sarker, S. Aridhi, J. Gomes, J. 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

[20] A. Moniot, I. Chauvot de Beauchêne and Y. Guermeur. 'Inferring Epsilon-nets of Finite Sets in a RKHS'. In: *Proceedings of the 14th International Workshop, Lecture Notes in Networks and Systems - LNNS*. Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization. WSOM - 2022. Vol. 533. Prague, Czech Republic: Springer, 27th Aug. 2022, pp. 53–62. DOI: 10.1007/978-3-031-15444-7_6. URL: https://hal.science/hal-03651323.

[21] J. Niehren, A. Vaginay and C. Versari. 'Abstract Simulation of Reaction Networks via Boolean Networks'. In: CMSB2022: 20th International Conference on Computational Methods in Systems Biology. Bucarest, Romania: Springer, 14th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-02279942.

Other scientific publications

[22] D. Amaya-Ramirez, R. Lhotte, M. Devriese, C. Hays, J.-L. Taupin and M.-D. Devignes. 'Reduced structural flexibility of eplet amino acids in HLA proteins'. In: ECCB 2022 21st European Conference on Computational Biology Planetary Health and Biodiversity. Barcelona, Spain, 12th Sept. 2022. URL: https://hal.inria.fr/hal-03924018.

[23] E. Bresso, P. Monnin, C. Bousquet, F.-É. 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*. Rennes, France, 5th July 2022. URL: https://hal.inria.fr/hal-03754898.

[24] H. Dhondge, I. Chauvot de Beauchêne and M.-D. Devignes. 'CroMaSt: A workflow for domain family curation through cross-mapping of structural instances between protein domain databases'. In: ECCB2022- 21st European Conference on Computational Biology. Sitges, Spain, 12th Sept. 2022. DOI: 10.48546/WORKFLOWHUB.WORKFLOW.390.1. URL: https://hal.archives-ouvertes.fr/hal-03789541.

12.3 Cited publications

[25] 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.

[26] 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.

[27] 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.

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

[29] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read and D. Baker. 'Accurate prediction of protein structures and interactions using a three-track neural network'. In: *Science* 373.6557 (Aug. 2021), pp. 871–876. DOI: 10.1126/science.abj8754.

[30] 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.

[31] D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao and A. Elofsson. 'Towards a structurally resolved human protein interaction network'. In: *Nat Struct Mol Biol* 30.2 (Feb. 2023), pp. 216–225.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] R. evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis. 'Protein complex prediction with AlphaFold-Multimer'. In: *bioRxiv* (2021). DOI: 10.1101/2021.10.04.463034. eprint: https://www.biorxiv.org/content/early/2021/10/04/2021.10.04.463034.full.pdf. URL: https://www.biorxiv.org/content/early/2021/10/04/2021.10.04.463034.

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

[38] 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.

[39] A. 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. DOI: 10.1093/nar/gkt1199. URL: https://hal.inria.fr/hal-00920612.

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

[41] A. 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.

[42] Y. Guermeur. 'L p -norm Sauer-Shelah Lemma for Margin Multi-category Classifiers'. In: *Journal of Computer and System Sciences* 89 (Nov. 2017), pp. 450–473. DOI: 10.1016/j.jcss.2017.06.003. URL: https://hal.science/hal-01371331.

[43] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian and S. E. Baranzini. 'Systematic integration of biomedical knowledge prioritizes drugs for repurposing'. In: *Elife* 6 (Sept. 2017).

[44] 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.

[45] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis. 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

[46] D. G. Kim, Y. Choi and H. S. Kim. 'Epitopes of Protein Binders Are Related to the Structural Flexibility of a Target Protein Surface'. In: *J Chem Inf Model* 61.4 (Apr. 2021), pp. 2099–2107.

[47] F. MacLean. 'Knowledge graphs and their applications in drug discovery'. In: *Expert Opin Drug Discov* 16.9 (Sept. 2021), pp. 1057–1069.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] D. J. Rigden and X. M. Fernández. 'The 2021 Nucleic Acids Research database issue and the online molecular biology database collection'. In: *Nucleic Acids Research* 49.D1 (Dec. 2020), pp. D1–D9. DOI: 10.1093/nar/gkaa1216. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D1/35364664/gkaa1216.pdf. URL: https://doi.org/10.1093/nar/gkaa1216.

[55] D. W. 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. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/17/2650/49021295/bioinformatics_32_17_2650.pdf. URL: https://doi.org/10.1093/bioinformatics/btw300.

[56] 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.

[57] 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.science/hal-02930827.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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