IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2018

Project-Team CAPSID

Computational Algorithms for Protein Structures and Interactions

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

RESEARCH CENTER
Nancy - Grand Est

THEME
Computational Biology

Table of contents

Project-Team CAPSID

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

Keywords:

Computer Science and Digital Science:

A1.5.1. - Systems of systems
A3.1.1. - Modeling, representation
A3.2.2. - Knowledge extraction, cleaning
A3.2.5. - Ontologies
A6.1.5. - Multiphysics modeling

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

1. Team, Visitors, External Collaborators

Research Scientists

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

Faculty Members

Sabeur Aridhi [Univ de Lorraine, Associate Professor]
Malika Smaïl-Tabbone [Univ de Lorraine, Associate Professor, from May 2018, HDR]

PhD Students

Seyed Ziaeddin Alborzi [Univ de Lorraine, until Aug 2018]
Kévin Dalleau [CNRS, from Dec 2016, ANR/Région Lorraine]
Antoine Moniot [Univ de Lorraine, from Oct 2018]
Gabin Personeni [CNRS, from Mar 2018 until Nov 2018]
Maria Elisa Ruiz Echartea [Inria, from Nov 2016, CORDI-S]
Bishnu Sarker [Inria, from Nov 2017, CORDI-S]
Athenais Vaginay [Univ de Lorraine, from Oct 2018]

Technical staff

Gabin Personeni [CNRS, from Nov 2018]
Emmanuel Bresso [CNRS]
Claire Lacomblez [CNRS, from Dec 2017]
Philippe Noël [Inria Apprentice, from Sep 2017 until Aug 2019]

Interns

Agnibha Chandra [Inria, from May 2018 until Jul 2018]
Ismail El Fadli [Univ de Lorraine, from Feb 2018 until Aug 2018]
Aichata Niang [Univ de Lorraine, from May 2018 until Jul 2018]
Rohit Roy [Inria, from May 2018 until Jul 2018]

Giammarco Mastronardi [Inria, from Feb 2018 until Jul 2018]
Wissem Inoubli [Univ de Lorraine, from Jun 2018 until Jul 2018]
Xavier Farchetto [Univ de Lorraine, from Jun 2018 until Aug 2018]
Maxime Guyot [CNRS, from Jun 2018 until Aug 2018]
Damien Vantourout [CNRS, from Jun 2018 until Jul 2018]

Administrative Assistants

Antoinette Courrier [CNRS]
Isabelle Herlich [Inria]
Annick Jacquot [CNRS, from Jul 2018]

Visiting Scientists

Patricia Alves Silva [Univ de Brasilia, from Oct 2018]
Ghania Khensous [Univ des Sciences et Technologies, Oran, Oct 2018]

External Collaborators

Taha Boukhobza [Univ de Lorraine, from Oct 2018]
Sjoerd Jacob de Vries [INSERM, from Sep 2018]
Vincent Leroux [Univ Denis Diderot, from Jun 2018]

2. Overall Objectives

2.1. Computational Challenges in Structural Biology

Many of the processes within living organisms can be studied and understood in terms of biochemical interactions between large macromolecules such as DNA, RNA, and proteins. To a first approximation, DNA may be considered to encode the blueprint for life, whereas proteins and RNA make up the three-dimensional (3D) molecular machinery. Many biological processes are governed by complex systems of proteins which interact cooperatively to regulate the chemical composition within a cell or to carry out a wide range of biochemical processes such as photosynthesis, metabolism, and cell signalling, for example. It is becoming increasingly feasible to isolate and characterise some of the individual protein components of such systems, but it still remains extremely difficult to achieve detailed models of how these complex systems actually work. Consequently, a new multidisciplinary approach called integrative structural biology has emerged which aims to bring together experimental data from a wide range of sources and resolution scales in order to meet this challenge [88], [67].

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

The overall objective of the Capsid team is to develop algorithms and software to help study biological systems and phenomena from a structural point of view. In particular, the team aims to develop algorithms which can help to model the structures of large multi-component biomolecular machines and to develop tools and techniques to represent and mine knowledge of the 3D shapes of proteins and protein-protein interactions. Thus, a unifying theme of the team is to tackle the recurring problem of representing and reasoning about large 3D macromolecular shapes. More specifically, our aim is to develop computational techniques to represent, analyse, and compare the shapes and interactions of protein molecules in order to help better understand how their 3D structures relate to their biological function. In summary, the Capsid team focuses on the following closely related topics in structural bioinformatics:

- new approaches for knowledge discovery in structural databases,
- integrative multi-component assembly and modeling.

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

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

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

3. Research Program

3.1. Classifying and Mining Protein Structures and Protein Interactions

3.1.1. Context

The scientific discovery process is very often based on cycles of measurement, classification, and generalisation. It is easy to argue that this is especially true in the biological sciences. The proteins that exist today represent the molecular product of some three billion years of evolution. Therefore, comparing protein sequences and structures is important for understanding their functional and evolutionary relationships [85], [57]. 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 [61], [90]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [42].

3.1.2. Quantifying Structural Similarity

Often, proteins may be divided into modular sub-units called domains, which can be associated with specific biological functions. Thus, a protein domain may be considered as the evolutionary unit of biological structure and function [89]. However, while it is well known that the 3D structures of protein domains are often more evolutionarily conserved than their one-dimensional (1D) amino acid sequences, comparing 3D structures is much more difficult than comparing 1D sequences. However, until recently, most evolutionary studies of proteins have compared and clustered 1D amino acid and nucleotide sequences rather than 3D molecular structures.

A pre-requisite for the accurate comparison of protein structures is to have a reliable method for quantifying the structural similarity between pairs of proteins. We recently developed a new protein structure alignment program called Kpax which combines an efficient dynamic programming based scoring function with a simple but novel Gaussian representation of protein backbone shape [76]. This means that we can now quantitatively compare 3D protein domains at a similar rate to throughput to conventional protein sequence comparison algorithms. We recently compared Kpax with a large number of other structure alignment programs, and we found Kpax to be the fastest and amongst the most accurate, in a CATH family recognition test [64]. The latest version of Kpax [9] can calculate multiple flexible alignments, and thus promises to avoid such issues when comparing more distantly related protein folds and fold families.

3.1.3. Formalising and Exploiting Domain Knowledge

Concerning protein structure classification, we aim to explore novel classification paradigms to circumvent the problems encountered with existing hierarchical classifications of protein folds and domains. In particular it will be interesting to set up fuzzy clustering methods taking advantage of our previous work on gene functional classification [50], 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) [55].

Another example where domain knowledge can be useful is during result interpretation: several sources of knowledge have to be used to explicitly characterise each cluster and to help decide its validity. Thus, it will be useful to be able to express data models, patterns, and rules in a common formalism using a defined vocabulary for concepts and relationships. Existing approaches such as the Molecular Interaction (MI) format [58] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [69]. A different point of view is represented in the Interaction Network Ontology (INO), a community-driven ontology that aims to standardise and integrate data on interaction networks and to support computer-assisted reasoning [92]. However, this ontology does not integrate basic 3D concepts and structural relationships. Therefore, extending such formalisms and symbolic relationships will be beneficial, if not essential, when classifying the 3D shapes of proteins at the domain family level.

3.1.4. 3D Protein Domain Annotation and Shape Mining

A widely used collection of protein domain families is “Pfam” [54], constructed from multiple alignments of protein sequences. Integrating domain-domain similarity measures with knowledge about domain binding sites, as introduced by us in our KBDOCK approach [2], [4], can help in selecting interesting subsets of domain pairs before clustering. Thanks to our KBDOCK and Kpax projects, we already have a rich set of tools with which we can start to process and compare all known protein structures and PPIs according to their component Pfam domains. Linking this new classification to the latest “SIFTS” (Structure Integration with Function, Taxonomy and Sequence) [86] functional annotations between standard UniProt (http://www.uniprot.org/ sequence identifiers and protein structures from the Protein Data Bank (PDB) [41] could then provide a useful way to discover new structural and functional relationships which are difficult to detect in existing classification schemes such as CATH or SCOP. As part of the thesis project of Seyed Alborzi, we developed a recommender-based data mining technique to associate enzyme classification code numbers with Pfam domains using our recently developed EC-DomainMiner program [1]. We subsequently generalised this approach as a tripartite graph mining method for inferring associations between different protein annotation sources, which we call “CODAC” (for COmputational Discovery of Direct Associations using Common Neighbours). A first paper on CODAC was presented at IWBBIO-2017 [36], and a full paper has recently been accepted by BMC Bioinformatics [13].

3.1.5. Protein Function Annotation

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 lack any functional annotation, and the annotation of sequences in such databases is a highly challenging problem. We are developing graph-based and machine learning techniques to annotate automatically the available unannotated sequences in such databases with functional properties such as EC numbers and Gene Ontology (GO) terms. Even if the 3D structures of proteins are unknown, it is natural to suppose that their sequences may be related to each other by the domains, domain families, and super-families that they share. In the frame of the PhD project of Bishnu Sarker, we recently developed a novel graph-based approach called GrAPFI for the automatic functional annotation of protein sequences based on these principles in order to transfer annotations from expert-reviewed sequences to unreviewed sequences in the UniProtKB databases [32], [24].

3.2. Integrative Multi-Component Assembly and Modeling

3.2.1. Context

At the molecular level, each PPI is embodied by a physical 3D protein-protein interface. Therefore, if the 3D structures of a pair of interacting proteins are known, it should in principle be possible for a docking algorithm to use this knowledge to predict the structure of the complex. However, modeling protein flexibility accurately during docking is very computationally expensive due to the very large number of internal degrees of freedom in each protein, associated with twisting motions around covalent bonds. Therefore, it is highly impractical to use detailed force-field or geometric representations in a brute-force docking search. Instead, most protein docking algorithms use fast heuristic methods to perform an initial rigid-body search in order to locate a relatively small number of candidate binding orientations, and these are then refined using a more expensive interaction potential or force-field model, which might also include flexible refinement using molecular dynamics (MD), for example.

3.2.2. Polar Fourier Docking Correlations

In our *Hex* protein docking program [77], 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 [60]

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

This equation demonstrates using the notion of *overlap* between 3D scalar quantities to give a physics-based scoring function. If the aim is to find the configuration that gives the most favourable interaction energy, then it is necessary to perform a six-dimensional search in the space of available rotational and translational degrees of freedom. By re-writing the polar Fourier expansions using complex spherical harmonics, we showed previously that fast Fourier transform (FFT) techniques may be used to accelerate the search in up to five of the six degrees of freedom [78]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [10], [6]. 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 [79], [62], it still remains challenging to produce a high resolution 3D model of a protein complex using *ab initio* techniques, mainly due to the problem of structural flexibility described above. However, with the aid of even just one simple constraint on the docking search space, the quality of docking predictions can improve considerably [10], [78]. In particular, many protein complexes involve symmetric arrangements of one or more sub-units, and the presence of symmetry may be exploited to reduce the search space considerably [40], [75], [84]. For example, using our operator notation (in which $\widehat{R}$ and $\widehat{T}$ represent 3D rotation and translation operators, respectively), we have developed an algorithm which can generate and score candidate docking orientations for monomers that assemble into cyclic (C_n) multimers using 3D integrals of the form

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

where the identical monomers A and B are initially placed at the origin, and $\omega_n = 2\pi/n$ is the rotation about the principal n-fold symmetry axis. This example shows that complexes with cyclic symmetry have just 4 rigid body degrees of freedom (DOFs), compared to $6(n-1)$ DOFs for non-symmetrical n-mers. We have generalised these ideas in order to model protein complexes that crystallise into any of the naturally occurring point group symmetries (C_n, D_n, T, O, I). This approach was published in 2016 [8], and was subsequently applied to several symmetrical complexes from the "CAPRI" blind docking experiment [53]. Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to build and refine candidate solutions using a more accurate coarse-grained (CG) force-field scoring function.

3.2.4. Coarse-Grained Models

Many approaches have been proposed in the literature to take into account protein flexibility during docking. The most thorough methods rely on expensive atomistic simulations using MD. However, much of a MD trajectory is unlikely to be relevant to a docking encounter unless it is constrained to explore a putative protein-protein interface. Consequently, MD is normally only used to refine a small number of candidate rigid body docking poses. A much faster, but more approximate method is to use 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 [68], [52], [65], [66]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [87], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [3].

In the last few years, CG *force-field* models have become increasingly popular in the MD community because they allow very large biomolecular systems to be simulated using conventional MD programs [39]. Typically, a CG force-field representation replaces the atoms in each amino acid with from 2 to 4 "pseudo-atoms", and it assigns each pseudo-atom a small number of parameters to represent its chemo-physical properties. By directly attacking the quadratic nature of pair-wise energy functions, coarse-graining can speed up MD simulations by up to three orders of magnitude. Nonetheless, such CG models can still produce useful models of very large multi-component assemblies [83]. Furthermore, this kind of coarse-graining effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [59]. We are therefore developing a "coarse-grained" scoring function for fast protein-protein docking and multi-component assembly in the frame of the PhD project of Maria-Elisa Ruiz-Echartea [31], [82].

3.2.5. Assembling Multi-Component Complexes and Integrative Structure Modeling

We also want to develop related approaches for integrative structure modeling using cryo-electron microscopy (cryo-EM). Thanks to recently 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 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 [49], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space (thesis project of Maria Elisa Ruiz Echartea).

4. Application Domains

4.1. Biomedical Knowledge Discovery

Participants: Marie-Dominique Devignes [contact person], Malika Smaïl-Tabbone, Sabeur Aridhi, David Ritchie, Gabin Personeni, Seyed Ziaeddin Alborzi, Kévin Dalleau, Bishnu Sarker, Emmanuel Bresso, Claire Lacomblez.

This projects in this domain are carried out in collaboration with the Orpailleur Team.

Huge and ever increasing amounts of biomedical data ("Big Data") are bringing new challenges and novel opportunities for knowledge discovery in biomedicine. We are actively collaborating with biologists and clinicians to design and implement approaches for selecting, integrating, and mining biomedical data in various areas. In particular, we are focusing on leveraging bio-ontologies at all steps of this process (the main thesis topic of Gabin Personeni, co-supervised by Marie-Dominique Devignes and Adrien Coulet from the Orpailleur team). One specific application concerns exploiting Linked Open Data (LOD) to characterise the genes responsible for intellectual deficiency. This work is in collaboration with Pr. P. Jonveaux of the Laboratoire de Génétique Humaine at CHRU Nancy [71], [72]. This involves using inductive logic programming as a machine learning method and at least three different ontologies (Gene Ontology, Human Phenotype Ontology, and Disease Ontology). This approach has also been applied using pattern structure mining (an extension of formal concept analysis) of drug and disease ontologies to discover frequently associated adverse drug events in patients [70]. This work was performed in collaboration with the Centre for BioMedical Informatics Research (BMIR) at Stanford University.

Recently, a new application for biomedical knowledge discovery has emerged from the ANR "FIGHT-HF" (fight heart failure) project, which is in collaboration with several INSERM teams at CHRU Nancy. In this case, the molecular mechanisms that underly HF at the cellular and tissue levels will be considered against a background of all available data and ontologies, and represented in a single integrated complex network. A network platform is under construction with the help of a young start-up company called Edgeleap. Together with this company, we are developing query and analysis facilities to help biologists and clinicians to identify relevant biomarkers for patient phenotyping [48]. Docking of small molecules on candidate receptors, as well as protein-protein docking will also be used to clarify a certain number of relations in the complex HF network.

4.2. Prokaryotic Type IV Secretion Systems

Participants: Marie-Dominique Devignes [contact person], Bernard Maigret, Isaure Chauvot de Beauchêne, David Ritchie, Philippe Noël, Aichata Niang.

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 [37]. 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 [56], [80]. However, the detailed nature of the interactions between the remaining components and those of the core channel remains to be resolved. Therefore, these secretion systems represent another family of complex biological systems (scales 2 and 3) that call for integrated modeling approaches to fully understand their machinery.

In the frame of the Lorraine Université d'Excellence (LUE-FEDER) "CITRAM" project MD Devignes is pursuing her collaboration with Nathalie Leblond of the Genome Dynamics and Microbial Adaptation (DynAMic) laboratory (UMR 1128, Université de Lorraine, INRA) on the discovery of new 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 genome. As these elements encode all or a subset of the components of the Type IV secretion system, they constitute a valuable source of sequence data and constraints for modeling these systems in 3D. Another interesting aspect of this particular system is that unlike other secretion systems, the Type IV secretion systems are not restricted to a particular group of bacteria [46].

4.3. Protein-RNA Interactions

Participants: Isaure Chauvot de Beauchêne [contact person], Antoine Moniot, Bernard Maigret, Maria Elisa Ruiz Echartea, David Ritchie, Agnibha Chandra, Rohit Roy.

As well as playing an essential role in the translation of DNA into proteins, RNA molecules carry out many other essential biological functions in cells, often through their interactions with proteins. A critical challenge in modelling such interactions computationally is that the RNA is often highly flexible, especially in single-stranded (ssRNA) regions of its structure. These flexible regions are often very important because it is through their flexibility that the RNA can adjust its 3D conformation in order to bind to a protein surface. However, conventional protein-protein docking algorithms generally assume that the 3D structures to be docked are rigid, and so are not suitable for modeling protein-RNA interactions. There is therefore much interest in developing protein-RNA docking algorithms which can take RNA flexibility into account.

We are currently developing a novel flexible docking algorithm which 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 [44], [45]. We have since implemented a prototype "forward-backward" dynamic programming algorithm with stochastic backtracking that allows us to model protein RNA interactions for ssRNAs of up to 7 nucleotides without requiring any prior knowledge of the interaction, while still avoiding a brute-force search. In the frame of our PEPS collaboration "InterANRIL" with the IMoPA lab (University of Lorraine), we are currently working with biologists to apply the approach to modeling certain long non-coding RNA (lncRNA) complexes. In order to extend this approach to partially structured RNA molecules, we have built an automated pipeline to create (i) libraries of RNA fragments with arbitrary characteristics such as secondary structure, and (ii) testing benchmarks for applying those libraries in docking. In the frame of our LUE-FEDER CITRAM project we adapted this approach and this pipeline to DNA docking in order to model the complex formed by a bacterial relaxase and its target DNA.

5. Highlights of the Year

5.1. Highlights of the Year

Isaure Chauvot de Beauchêne has obtained H2020 funding for two international PhD students under the MSCA-ITN programme. The project will study protein/RNA interactions, and will start on 01/01/2019.

6. New Software and Platforms

6.1. Hex

KEYWORDS: 3D rendering - Bioinformatics - 3D interaction - Structural Biology

SCIENTIFIC DESCRIPTION: Hex is an interactive protein docking and molecular superposition program for Linux Mac-OS and Windows-XP. Hex understands protein and DNA structures in PDB format, and it can also read small-molecule SDF files. The recent versions now include CUDA support for Nvidia GPUs. On a modern workstation, docking times range from a few minutes or less when the search is constrained to known binding sites, to about half an hour for a blind global search (or just a few seconds with CUDA).
FUNCTIONAL DESCRIPTION: The underlying algorithm uses a novel polar Fourier correlation technique to accelerate the search for close-fitting orientations of the two molecules.

- Participant: David Ritchie
- Contact: David Ritchie
- URL: http://hex.loria.fr

6.2. Kbdock

KEYWORD: 3D interaction
SCIENTIFIC DESCRIPTION: Kbdock is a database of 3D protein domain-domain interactions with a web interface.
FUNCTIONAL DESCRIPTION: The Kbdock database is built from a snapshot of the Protein Databank (PDB) in which all 3D structures are cut into domains according to the Pfam domain description. A web interface allows 3D domain-domain interactions to be compared by Pfam family.

- Authors: Anisah Ghoorah, David Ritchie and Marie-Dominique Devignes
- Contact: David Ritchie
- URL: http://kbdock.loria.fr

6.3. Kpax

KEYWORDS: Bioinformatics - Structural Biology
SCIENTIFIC DESCRIPTION: Kpax is a program for aligning and superposing the 3D structures of protein molecules.
FUNCTIONAL DESCRIPTION: The algorithm uses a Gaussian representation of the protein backbone in order to construct a similarity score based on the 3D overlap of the Gaussians of the proteins to be superposed. Multiple proteins may be aligned together (multiple structural alignment) and databases of protein structures may be searched rapidly.

- Participant: David Ritchie
- Contact: David Ritchie

6.4. Sam

Protein Symmetry Assembler
KEYWORDS: Proteins - Structural Biology
SCIENTIFIC DESCRIPTION: Sam is a program for making symmetrical protein complexes, starting from a single monomer.
FUNCTIONAL DESCRIPTION: The algorithm searches for good docking solutions between protein monomers using a spherical polar Fast Fourier transform correlation in which symmetry restraints are built into the calculation. Thus every candidate solution is guaranteed to have the desired symmetry.

- Authors: David Ritchie and Sergey Grudinin
- Partner: CNRS
- Contact: David Ritchie
- URL: http://sam.loria.fr

6.5. gEMfitter

KEYWORDS: 3D reconstruction - Cryo-electron microscopy - Fitting
SCIENTIFIC DESCRIPTION: A program for fitting high resolution 3D protein structures into low resolution cryo-EM density maps.
FUNCTIONAL DESCRIPTION: A highly parallel fast Fourier transform (FFT) EM density fitting program which can exploit the special hardware properties of modern graphics processor units (GPUs) to accelerate both the translational and rotational parts of the correlation search.

- Authors: Van-Thai Hoang and David Ritchie
- Contact: David Ritchie
- URL: http://gem.loria.fr/gEMfitter/

6.6. ECDM

ECDomainMiner
KEYWORD: Functional annotation
SCIENTIFIC DESCRIPTION: EC-DomainMiner uses a recommender-based approach for associating EC (Enzyme Commission) numbers with protein Pfam domains from EC-sequence relationships that have been annotated previously in the SIFTS and Uniprot databases.
FUNCTIONAL DESCRIPTION: A program to associate protein Enzyme Commission numbers with Pfam domains

- Contact: David Ritchie
- URL: http://ecdm.loria.fr

6.7. GODM

GO-DomainMiner
KEYWORD: Functional annotation
FUNCTIONAL DESCRIPTION: GO-DomainMiner is is a graph-based approach for associating GO (gene ontology) terms with protein Pfam domains.

- Contact: David Ritchie
- URL: http://godm.loria.fr

6.8. BLADYG

A Block-centric graph processing framework for LArge Dynamic Graphs
KEYWORDS: Distributed computing - Dynamic graph processing
FUNCTIONAL DESCRIPTION: BLADYG is a block-centric framework that addresses the issue of dynamism in large-scale graphs. BLADYG starts its computation by collecting the graph data from various data sources. After collecting the graph data, BLADYG partitions the input graph into multiple partitions. Each BLADYG worker loads its block/partition and performs both local and remote computations, after which the status of the blocks is updated. The BLADYG coordinator orchestrates the execution of the considered graph operation in order to deal with graph updates.

- Partner: University of Trento
- Contact: Sabeur Aridhi

6.9. CGC

Clebsch-Gordan Coefficients
KEYWORDS: Clebsch-Gordan coupling coefficient - 3j symbol

FUNCTIONAL DESCRIPTION: Clebsch-Gordan coupling coefficients appear in many areas of physics and chemistry. CGC is a small library of functions and a demo driver program for calculating Clebsch-Gordan coupling coefficients up to very high principal quantum numbers.

- Contact: David Ritchie
- URL: http://cgc.loria.fr

6.10. GrAPFI

GrAPFI: Graph-based Automatic Protein Function Inference
KEYWORD: Proteins
FUNCTIONAL DESCRIPTION: GrAPFI is a Graph-based Automatic Protein Function Inference tool that aims to annotate protein sequences with EC numbers. The underlying philosophy of GrAPFI assumes that proteins can be linked through the domains, families, and superfamilies that they share. Several domain databases exist such as e.g. Pfam, SMART, CDD, Gene3D, and Prosite. Furthermore, InterPro aims to integrate information from all such databases by assigning them unique InterPro signatures. GrAPFI tool also shares Interpro signatures, as it includes information from several major family and domain databases. Our computational analysis and cross-validation show that GrAPFI achieves state-of-the-art performance in EC number prediction.

- Contact: Sabeur Aridhi
- URL: http://grapfi.loria.fr/

6.11. Platforms

6.11.1. The MBI Platform

The MBI (Modeling Biomolecular Interactions) platform (http://bioinfo.loria.fr) was established to support collaborations between Inria Nancy – Grand Est and other research teams associated with the University of Lorraine. The platform is a research node of the Institut Français de Bioinformatique (IFB), which is the French national network of bioinformatics platforms (http://www.france-bioinformatique.fr). **In 2018, funding for an engineer was awarded to Marie-Dominique Devignes for a project on bioinformatics service integration.**

- Contact: Marie-Dominique Devignes

7. New Results

7.1. Drug Targeting and Adverse Drug Side Effects

Identifying new molecular targets using comparative genomics and knowledge of disease mechanisms is a rational first step in the search for new preventative or therapeutic drug treatments [63]. We are mostly concerned with three global health problems, namely fungal and bacterial infections and hypertension. Through on-going collaborations with several Brazilian laboratories (at University of Mato Grosso State, University of Maringá, Embrapa, and University of Brasilia), we previously identified several novel small-molecule drug leads against *Trypanosoma cruzi*, a parasite responsible for Chagas disease [91]. With the University of Maringá, we subsequently found several active molecules against the flavoenzyme TRR1 in *Candida albicans*, and two manuscripts are in preparation. We also proposed several small-molecule inhibitors against *Fusarium graminearum*, a fungal threat to global wheat production [63], [43]. Two further manuscripts on this topic are currently in preparation. Concerning hypertension, we continued our collaboration with Prof. Catherine Llorens-Cortes at Collège de France to study the interaction between the apelin receptor (a transmembrane protein important for blood pressure regulation) and the aminopetidase A enzyme [47].

It is well known that many therapeutic drug molecules can have adverse side effects. However, when patients take several combinations of drugs it can be difficult to determine which drug is responsible for which side effect. In collaboration with Adrien Coulet (Orpailleur team co-supervisor of Gabin Personeni) and Prof. Michel Dumontier (Biomedical Informatics Research Laboratory, Stanford), we developed an approach which combines multiple ontologies such as the Anatomical Therapeutical Classification of Drugs, the ICD-9 classification of diseases, and the SNOMED-CT medical vocabulary together with the use of Pattern Structures (an extension of Formal Concept Analysis) in order to extract association rules to analyse the co-occurrence of adverse drug effects in patient records [74], [73]. A paper describing this work has been published in the Journal of Biomedical Semantics [70].

7.2. Docking Symmetrical Protein Structures

Many proteins form symmetrical complexes in which each structure contains two or more identical copies of the same sub-unit. We recently developed a novel polar Fourier docking algorithm called "Sam" for automatically assembling symmetrical protein complexes. A journal article describing the Sam algorithm has been published [8]. An article describing the results obtained when using Sam to dock several symmetrical protein complexes from the "CASP/CAPRI" docking experiment has also been published [53]. This study showed that many of the models of protein structures built by members of the "CASP" fold prediction community are "dockable" in the sense that Sam is able to find acceptable docking solutions from amongst the CASP models.

More recently, we are working to extend the polar Fourier correlation algorithm to use very high angular resolution spherical Bessel basis functions. As part of this work, we have developed a very fast recursive algorithm for calculating high order Clebsch-Gordan coupling coefficients [30]. A manuscript describing this work has been submitted to a quantum mechanics journal.

7.3. Multiple Flexible Protein Structure Alignments

Comparing two or more proteins by optimally aligning and superposing their backbone structures provides a way to detect evolutionary relationships between proteins that cannot be detected by comparing only their primary amino-acid sequences. The latest version of our "Kpax" protein structure alignment algorithm can flexibly align pairs of structures that cannot be completely superposed by a single rigid-body transformation, and can calculate multiple alignments of several similar structures flexibly [9]. In collaboration with Alain Hein of the INRA lab "Agronomie et Environnement", we used Kpax to help study the structures of various "Cyp450" enzymes in plants [81]. In collaboration with Emmanuel Levy of the Weizmann Institute, we used Kpax to superpose and compare all of the symmetrical protein complexes in the Protein Databank in order to verify or remediate their quaternary structure annotations. A manuscript describing this work has been published in Nature Methods [15].

7.4. Large-Scale Annotation of Protein Domains and Sequences

Many protein chains in the Protein Data Bank (PDB) are cross-referenced with Pfam domains and Gene Ontology (GO) terms. However, these annotations do not explicitly indicate any relation between EC numbers and Pfam domains, and many others lack GO annotations. In order to address this limitation, as part of the PhD thesis project of Seyed Alborzi, we developed the CODAC approach for mining multiple protein data sources (i.e. SwissProt, TremBL, and SIFTS) in order to associate GO molecular function terms with Pfam domains, for example. We named the software implementation "GO-DomainMiner". This work was first presented at IWBBIO 2017 [36]. A full paper has recently been accepted for a special issue of *BMC Bioinformatics* [13].

In collaboration with Maria Martin's team at the European Bioinformatics Institute (EBI), we combined the CODAC approach with a novel combinatorial association rule based approach called "CARDM" for annotating protein sequences. When applied to the large UniProt/TrEMBL sequence database of 63 million protein entries, CARDM predicted over 24 million Enzyme Commission (EC) numbers and 188 million GO terms for those entries. A journal paper in collaboration with the EBI on comparing the quality of these

predicted annotations with other state of the art annotation methods is in preparation, and a poster was presented at ISMB-ECCB-2017 [35]. As part of the PhD thesis of Bishnu Sarker, we also developed GrAPFI, a graph-based protein function annotation approach. GrAPFI applies a label propagation algorithm to a complex network representation of protein sequence data. A full paper on this work has recently been accepted by the International Conference on Complex Networks and their Applications [24].

7.5. Distributed Protein Graph Processing

The huge number of protein sequences in protein databases such as UniProtKB calls for rapid procedures to annotate them automatically. We are using existing protein annotations to predict the annotations of new or non-reviewed proteins. In this context, we developed the "DistNBLP" method for annotating protein sequences using a graph representation and a distributed label propagation algorithm. DistNBLP uses the BLADYG framework [38] to process protein graphs on multiple compute nodes by applying a neighbourhood-based label propagation algorithm in a distributed way. We applied DistNBLP in the recent "CAFA 3" (critical Assessment of Protein Function Annotation) community experiment to annotate new protein sequences automatically. This work was presented as a poster at ISMB/ECCB-2017 [34]. We are also interested in feature selection for subgraph patterns. In collaboration with the LIMOS laboratory at Université Clermont Auvergne we also developed a scalable approach using MapReduce for identifying sub-graphs having similar labels in very large graphs [51].

7.6. Flexible Docking of Protein-GAG Complexes

Modeling how flexible polymers bind to proteins presents enormous computational challenges due to the large conformational search space that arises from the many internal rotational degrees of freedom in polymer structures. In collaboration with Sergey Samsonov (Gdansk University, Poland), we extended our fragment-based flexible docking approach [83], [42] to model how flexible Glycosaminoglycans (GAGs) might bind to the surface of a known protein structure. A paper has been submitted to the Journal of Computational Chemistry.

In collaboration with Sjoerd de Vries (Univ Paris Diderot), we have created a new protein-glycan interaction force-field and integrated it in the ATTRACT docking engine [83]. We also participated in a comparative study of the main current protein-GAG docking methods.

7.7. Stochastic Decision Trees for Similarity Computation

We have designed a method to compute similarities on unlabeled data using stochastic decision trees [20]. The main idea of Unsupervised Extremely Randomized Trees (UET) is to randomly and iteratively split the data until a stopping criterion is met. Pairwise similarity values are computed based on the co-occurrence of samples in the leaves of each generated tree. We evaluate our method on synthetic and real-world datasets by comparing the mean similarities between samples with the same label and the mean similarities between samples with distinct labels. Empirical studies show that the method effectively gives distinct similarity values between samples belonging to distinct clusters, and gives indiscernible values when there is no cluster structure. We also assessed some interesting properties such as invariance under monotone transformations of variables and robustness to correlated variables and noise. Our experiments show that the algorithm outperforms existing methods in some cases, and can reduce the amount of preprocessing needed with many real-world datasets. We plan to study the application of this "global" pairwise similarity computation to quantify protein structural similarities. Two interesting problems will concern the representation of the protein structure and how to tackle extra constraints such as invariance under rotational and translational transformations.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. CPER – IT2MP

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

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

8.1.2. LUE-FEDER – CITRAM

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

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

8.1.3. PEPS – DynaCriGalT

Participants: Isaure Chauvot de Beauchêne [contact person], Bernard Maigret, David Ritchie.

Project title: *Criblage virtuel et dynamique moléculaire pour la recherche de bio-actifs ciblant la β4GalT7, une enzyme de biosynthèse des glycosaminoglycanes*; PI: I. Chauvot de Beauchêne, Capsid (Inria Nancy – Grand Est); Partners: Sylvie Fournel-Gigleux, INSERM (IMoPA, UMR 7365); Value: 15 k€; Duration: 2017–2018. Description: The β4GalT7 glycosyltransferase initiates the biosynthesis of glycosaminoglycans (GAGs), and is a therapeutical target for small molecules which might correct a defect in the synthesis and degradation of GAGs in rare genetic diseases. Classical approaches to propose active molecules have failed for this target. The DynaCriGalT project combines molecular dynamics modelling of the GAG active site with virtual screening in order to propose a diverse set of small molecules for *in vitro* compound testing.

8.1.4. PEPS – InterANRIL

Participant: Isaure Chauvot de Beauchêne [contact person].

Project title: *TBA* Duration: 2017–2018. Description: *TBA*

Project title: *Identification et modélisation des interactions nécessaires à l'activité du long ARN non-codant ANRIL dans la régulation épigénétique des gènes*; PI: Sylvain Maenner, Univ Lorraine (IMoPA, UMR 7365); Value: 20 k€; Duration: 2017–2018. Description: ANRIL is a long non-coding RNA (lncRNA) which has been identified as an important factor in the susceptibility cardiovascular diseases. ANRIL is involved in the epigenetic regulation of the expression of a network of genes via mechanisms that are still largely unknown. This project aims to identify and model the protein-RNA and/or DNA-RNA interactions that ANRIL establishes within the eukaryotic genome.

8.1.5. GlycoEst

Participant: Isaure Chauvot de Beauchêne [contact person].

GlycoEst is an informal working group which was recently created to develop an interdisciplinary regional network of Glyco-scientists. Isaure Chauvot de Beauchêne gave a talk on her protein-GAG docking method at the first meeting of this group in March 2018.

8.2. National Initiatives

8.2.1. FEDER – SB-Server

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

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

8.2.2. ANR

8.2.2.1. Fight-HF

Participants: Marie-Dominique Devignes [contact person], Malika Smaïl-Tabbone [contact person], Bernard Maigret, Sabeur Aridhi, Kévin Dalleau, Claire Lacomblez, David Ritchie.

Project title: *Combattre l'insuffisance cardiaque*; PI: Patrick Rossignol, Univ Lorraine (FHU-Cartage); Partners: multiple; Value: 9 m€ (Capsid and Orpailleur: 450 k€, approx); Duration: 2015–2019. Description: This "Investissements d'Avenir" project aims to discover novel mechanisms for heart failure and to propose decision support for precision medicine. The project has been granted € 9M, and involves many participants from Nancy University Hospital's Federation "CARTAGE" (http://www.fhu-cartage.com/). In collaboration with the Orpailleur Team, Marie-Dominique Devignes is coordinating a work-package on network-based science and drug discovery for this project.

8.2.2.2. IFB

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

Project title: *Institut Français de Bioinformatique*; PI: Claudine Médigue and Jacques van Helden (CNRS UMS 3601); Partners: multiple; Value: 20 M€ (Capsid: 126 k€); Duration: 2014–2021. Description: The Capsid team is a research node of the IFB (Institut Français de Bioinformatique), the French national network of bioinformatics platforms (http://www.france-bioinformatique.fr). The principal aim is to make bioinformatics skills and resources more accessible to French biology laboratories.

8.2.3. Collaborations with Major European Organizations

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

8.2.4. PEPS-INS2I – ORCA 3D

Participants: Isaure Chauvot de Beauchêne [contact person], Agnibha Chandra, Rohit Roy.

Protect Title: *Oligo-RNA Combinatorial Assembly for 3D modeling of protein-RNA complexes.* PI: Isaure Chauvot de Beauchêne. Value: 8k€. Description: The project aimed at improving our fragment-based ssRNA docking method, of which we already provided a proof of principle. It mainly provided grants for two internship students to work on (i) a new ssRNA-protein scoring function and (ii) docking with constraints specific to the geometry of ssRNA in RNA loops.

8.3. International Initiatives

8.3.1. TempoGraphs

Participants: Sabeur Aridhi [contact person], Marie-Dominique Devignes, Malika Smaïl-Tabbone, David Ritchie, Bishnu Sarker, Wissem Inoubli.

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

8.3.1.1. Informal International Partners

Participant: David Ritchie; Project: *Integrative Modeling of 3D Protein Structures and Interactions;* Partner: Rocasolano Institute of Physical Chemistry, Spain. Funding: Inria Nancy – Grand Est ("Nancy Emerging Associate Team").

Participant: Bernard Maigret; Project: *Characterization, expression and molecular modeling of TRR1 and ALS3 proteins of Candida spp., as a strategy to obtain new drugs with action on yeasts involved in nosocomial infections;* Partner: State University of Maringá, Brazil.

Participant: Bernard Maigret; Project: *Fusarium graminearum target selection;* Partner: Embrapa Recursos Geneticos e Biotecnologia, Brazil.

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

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

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Ghania Khensous from the University of Sciences and Technologies in Oman visited the team to develop a tabu-based search algorithm for flexible protein-ligand docking, under the supervision of Bernard Maigret.

Patricia Alves from the University of Brasilia is visiting the team to carry out drug repositioning on several target fungus proteins under the supervision of Bernard Maigret.

8.4.1.1. Internships

Agnibha Chandra from the Indian Institute of Engineering Science & Technology visited the team to optimize a force-field for ssRNA-protein docking, under the supervision of Isaure Chauvot de Beauchêne.

Ismail El Fadli from the Mohammed V University of Rabat visited the team to adapt our RNA-protein docking method to DNA-protein systems, under the supervision of Isaure Chauvot de Beauchêne.

Aichata Niang from the University of Paris Diderot visited the team to to apply modeling and virtual screening of a galactosyl-transferase enzyme in order to find new inhibitors, under the supervision of Isaure Chauvot de Beauchêne.

Rohit Roy from the Indian Institute of Technology at Kharagpur visited the team in order to include geometric constrains in our docking algorithm for docking RNA loops, under the supervision of Isaure Chauvot de Beauchêne.

Giammarco Mastronardi from the University of Lorraine visited the team to carry out a virtual screening study of small-molecule inhibitors of a bacterial polyketide synthase module, under the supervision of Bernard Maigret and David Ritchie.

Wissem Inoubli from the University of Tunis El Manar visited the team to work on his PhD thesis on distributed graph processing under the supervision of Sabeur Aridhi.

Damien Vantourout from the University of Lorraine visited the team to develop a tool for protein function annotation using semantic protein networks and deep neural networks under the supervision of Sabeur Aridhi.

Xavier Farchetto from the University of Lorraine visited the team to work on the segmentation of images in Crohn's disease under the supervision of Malika Smaïl-Tabbone and Chedy Raissy (Orpailleur team).

Maxime Guyot from the University of Lorraine (Telecom Nancy stage 2A) visited the team to perform statistical analyses on KBDock and to develop a scoring function for protein-protein interaction subgraphs extracted from a knowledge graph database.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Marie-Dominique Devignes is a member of the Steering Committee for the European Conference on Computational Biology (ECCB; http://eccb.iscb.org/). She was co-organizer of the national Workshop on AI and Health in the frame of PFIA 2018 (https://pfia2018.loria.fr/journeeiasante/), and she co-organized the Journée Entreprises (Atelier Santé) of the Fédération Charles Hermite (http://www.loria.fr/wp-content/uploads/2017/12/Programme-Forum-FCH-Entreprises.pdf).

Malika Smaïl-Tabbone is a member of the steering committee of the Conférence Francophone sur la Recherche d'Information et ses Applications (CORIA).

9.1.2. Scientific Events Selection

9.1.2.1. Member of the Conference Program Committees

Malika Smaïl-Tabbone is a member of the organizing committee for Inforsid 2018 and EGC 2018, and is in charge of the é-EGC winter school in Metz.

9.1.2.2. Reviewer

Marie-Dominique Devignes was a reviewer for IWBBIO and BIBM.

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

David Ritchie is a member of the editorial board of Scientific Reports.

Sabeur Aridhi is a member of the editorial board of Intelligent Data Analysis.

9.1.3.2. Reviewer - Reviewing Activities

The members of the team have reviewed manuscripts for *Bioinformatics, BMC Bioinformatics, Computational and Structural Biotechnology Journal, Computational Biology and Chemistry, Evolutionary Bioinformatics, Journal of Computational Chemistry, Journal of Chemical Information and Modeling, Journal of Molecular Graphics and Modeling, Nucleic Acids Research, PLoS One, Proteins: Structure, Function & Bioinformatics,* and *Structure.*

9.1.4. Invited Talks

Marie-Dominique Devignes gave a presentation on accelerating precision medicine to the OLINK proteomics society in Uppsala, Sweden.

Malika Smaïl-Tabbone gave a talk on "Integrative Machine Learning Applied on Drug Side Effect Profiles" at WCP2018 (18th World Congress of Basic and Clinical Pharmacology) in Kyoto, Japan.

Isaure Chauvot de Beauchêne gave a presentation on docking by combinatorial assembly of fragments to the Centre de mathématiques et de leurs applications (CMLA) at ENS Cachan.

9.1.5. Scientific Expertise

Marie-Dominique Devignes reviewed grant applications for the ANR programme "Appel générique-JCJC".

Malika Smaïl-Tabbone reviewed grant applications for the ANR and for the Indo-French Centre for the Promotion of Advanced Research.

Sabeur Aridhi reviewed grant applications for the French Committee for the Evaluation of Academic and Scientific Cooperation with Brazil (COFECUB).

David Ritchie is a member of the Bureau of the GGMM (Groupe de Graphisme et Modélisation Moléculaire).

Isaure Chauvot de Beauchêne is the team's representative to the ELIXIR 3D-Bioinfo working group, which is an essential link between the national IFB and the European ELIXIR projects that aim to make bioinformatics software and platforms available to non-expert biologists.

9.1.6. Research Administration

Marie-Dominique Devignes is Chargée de Mission for the CyberBioHealth research axis at the LORIA and is a member of the "Comipers" recruitment committee for Inria Nancy – Grand Est.

David Ritchie is a member of the Commission de Mention Informatique (CMI) of the University of Lorraine's IAEM doctoral school. Until November 2018 he was a member of the Bureau of the Project Committee for Inria Nancy – Grand Est.

Sabeur Aridhi is responsible for the major in IAMD (Ingénierie et Applications des Masses de Données) at TELECOM Nancy (Univ. Lorraine), and a member of the "Commission du Développement Technologique" recruitment committee at Inria Nancy – Grand Est.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence: Sabeur Aridhi, *P*rogramming Techniques and Tools, 24 hours, L1, Univ Lorraine.

Licence: Sabeur Aridhi, *D*atabases, 82 hours, L1, Univ Lorraine.

Licence: Sabeur Aridhi, *M*assive Data Management, 68 hours, L2, Univ Lorraine.

Licence: Sabeur Aridhi, *N*oSQL Databases, 44 hours, L2, Univ Lorraine.

Licence: Sabeur Aridhi, *B*ig Data Hackathon, 8 hours, L3, Univ Lorraine.

Licence: Marie-Dominique Devignes, *R*elational Database Design and SQL, 30 hours, L3, Univ Lorraine.

Licence: Isaure Chauvot de Beauchêne, *T*D Bioinformatique et Modelisation, 10 hours, L3, Univ Lorraine.

Licence: Malika Smaïl-Tabbone, *R*elational Databases, 90 hours, L2, L3, Univ Lorraine.

Licence: Malika Smaïl-Tabbone, *N*oSQL Databases, 30 hours, M1, Univ Lorraine.

Licence: Malika Smaïl-Tabbone, *P*rogramming Techniques, 30 hours, L2, Univ Lorraine.

Master: Malika Smaïl-Tabbone, *K*DD and Data Mining Algorithms, 90 hours, M2, Univ Lorraine.

Master: Malika Smaïl-Tabbone, *D*atabases : Concepts and Techniques, 30 hours, M2, Univ Lorraine.

Master: Malika Smaïl-Tabbone, *O*ntology Management and Semantic Web Technologies, 30 hours, M2, Univ Lorraine.

Master: Sabeur Aridhi, *K*nowledge Discovery and Data Engineering, 10 hours, M2, Univ Lorraine.

9.2.2. Supervision

PhD in progress: Maria Elisa Ruiz Echartea, *M*ulti-component protein assembly using distance constraints, 01/11/2016, David Ritchie, Isaure Chauvot de Beauchêne.

PhD in progress: Kévin Dalleau, *C*omplex graph analysis for classification: application to disease nosography, 01/12/2016, Malika Smaïl-Tabbone, Miguel Couerceiro.

PhD in progress: Bishnu Sarker, *D*eveloping distributed graph-based approaches for large-scale protein function annotation and knowledge discovery, 01/11/2017, David Ritchie, Sabeur Aridhi.

PhD in progress: Antoine Moniot, *M*odeling protein / nucleic acid complexes by combinatorial structural fragment assembly, 01/11/2017, David Ritchie, Isaure Chauvot de Beauchêne.

PhD in progress: Athénais Vaginay, *M*odel selection and analysis for biological networks: use of domain knowledge and application to networks disturbed in diseases,

01/11/2017, Taha Boukhobza, Malika Smaïl-Tabbone.

10. Bibliography

Major publications by the team in recent years

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

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

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

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

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

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

[7] V. PÉREZ-NUENO, A. S. KARABOGA, M. SOUCHET, D. RITCHIE. *GESSE: Predicting Drug Side Effects from Drug–Target Relationships*, in "Journal of Chemical Information and Modeling", August 2015, vol. 55, n° 9, pp. 1804-1823 [*DOI :* 10.1021/ACS.JCIM.5B00120], https://hal.inria.fr/hal-01216493

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

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

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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] S. Z. ALBORZI. *Automatic Discovery of Hidden Associations Using Vector Similarity : Application to Biological Annotation Prediction*, Université de Lorraine, February 2018, https://tel.archives-ouvertes.fr/tel-01792299

[12] G. PERSONENI. *Contribution of domain ontologies for knowledge discovery in biomedical data*, Université de Lorraine, November 2018, https://hal.inria.fr/tel-01925461

Articles in International Peer-Reviewed Journals

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

[14] I. R. G. CAPOSI, D. R. FARIA, K. M. SEIKETA, F. RODRIGUES, P. BOMFIM-MENDONÇA, T. BECKER, E. S. KIOSHIMA, T. SVIDZINSKI, B. MAIGRET. *Repurposing Approach Identifies New Treatment Options for Invasive Fungal Disease*, in "Bioorganic Chemistry", 2018 [*DOI :* 10.1016/J.BIOORG.2018.11.019], https://hal.inria.fr/hal-01927136

[15] S. DEY, D. RITCHIE, E. D. LEVY. *PDB-wide identification of biological assemblies from conserved quaternary structure geometry*, in "Nature Methods", 2018, vol. 15, pp. 67-72 [*DOI :* 10.1038/NMETH.4510], https://hal.inria.fr/hal-01652359

[16] K. GHANIA, B. MESSABIH, A. CHOUARFIA, B. MAIGRET. *Flexible molecular docking: application of hybrid tabu-simplex optimisation*, in "International journal of computational biology and drug design", 2018, https://hal.inria.fr/hal-01927105

[17] W. INOUBLI, S. ARIDHI, H. MEZNI, M. MADDOURI, E. MEPHU NGUIFO. *An experimental survey on big data frameworks*, in "Future Generation Computer Systems", September 2018, vol. 86, pp. 546 - 564 [*DOI :* 10.1016/J.FUTURE.2018.04.032], https://hal.inria.fr/hal-01926259

[18] C. KRIEGER, S. ROSELLI, S. KELLNER-THIELMANN, G. GALATI, B. SCHNEIDER, J. GROSJEAN, A. OLRY, D. RITCHIE, U. MATERN, F. BOURGAUD, A. HEHN. *The CYP71AZ P450 Subfamily: A Driving Factor for the Diversification of Coumarin Biosynthesis in Apiaceous Plants*, in "Frontiers in Plant Science", June 2018, vol. 9 [*DOI :* 10.3389/FPLS.2018.00820], https://hal.archives-ouvertes.fr/hal-01899038

[19] H. MEZNI, S. ARIDHI, A. HADJALI. *The uncertain cloud: State of the art and research challenges*, in "International Journal of Approximate Reasoning", December 2018, vol. 103, pp. 139 - 151 [*DOI :* 10.1016/J.IJAR.2018.09.009], https://hal.inria.fr/hal-01926257

International Conferences with Proceedings

[20] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Unsupervised extremely randomized trees*, in "PAKDD 2018 - The 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining", Melbourne, Australia, May 2018, https://hal.inria.fr/hal-01667317

[21] M.-D. DEVIGNES, Y. FRANSOT, Y. LEPAGE, J. LIEBER, E. NAUER, M. SMAÏL-TABBONE. *First steps toward finding relevant pathology-gene pairs using analogy*, in "EvoCBR 2018 : Workshop on Evolutionary Computation and CBR at the International Conference on Case-Based Reasoning (ICCBR 2018)", Stockholm, Sweden, July 2018, https://hal.inria.fr/hal-01906547

[22] W. INOUBLI, S. ARIDHI, H. MEZNI, M. MADDOURI, E. M. NGUIFO. *A Comparative Study on Streaming Frameworks for Big Data*, in "VLDB 2018 - 44th International Conference on Very Large Data Bases : Workshop LADaS - Latin American Data Science", Rio de Janeiro, Brazil, August 2018, pp. 1-8, https://hal.inria.fr/hal-01835437

[23] G. PERSONENI, M.-D. DEVIGNES, M. SMAÏL-TABBONE, P. JONVEAUX, C. BONNET, A. COULET. *Cooperation of bio-ontologies for the classification of genetic intellectual disabilities : a diseasome approach*, in "Proceedings of the 11th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS 2018)", Antwerp, Belgium, December 2018, https://hal.inria.fr/hal-01925471

[24] B. SARKER, D. W. RITCHIE, S. ARIDHI. *Exploiting Complex Protein Domain Networks for Protein Function Annotation*, in "Complex Networks 2018 - 7th International Conference on Complex Networks and Their Applications", Cambridge, United Kingdom, December 2018, https://hal.inria.fr/hal-01920595

National Conferences with Proceedings

[25] M. ZOGHLAMI, S. ARIDHI, M. MADDOURI, E. M. NGUIFO. *ABClass : Une approche d'apprentissage multi-instances pour les séquences*, in "RJCIA 2018 - 16èmes Rencontres des Jeunes Chercheurs en Intelligence Artificielle", Nancy, France, July 2018, pp. 1-9, https://hal.inria.fr/hal-01835432

Scientific Books (or Scientific Book chapters)

[26] M. ZOGHLAMI, S. ARIDHI, M. MADDOURI, E. MEPHU NGUIFO. *An Overview of in Silico Methods for the Prediction of Ionizing Radiation Resistance in Bacteria*, in "Ionizing Radiation: Advances in Research and Applications", T. REEVE (editor), Physics Research and Technology Series, Nova science publishers, May 2018, pp. 241-256, https://hal.inria.fr/hal-01807944

Other Publications

[27] S. Z. ALBORZI, S. ARIDHI, D. RITCHIE, M.-D. DEVIGNES. *PPI DomainMiner: predicting domain-domain interactions from protein-protein interactions using tripartite graph modeling and vector similarity*, September 2018, ECCB 2018 - 17th European Conference on Computational Biology, Poster, https://hal.inria.fr/hal-01877112

[28] E. BRESSO, C. LACOMBLEZ, A. PIZARD, P. ROSSIGNOL, F. ZANNAD, M. SMAÏL-TABBONE, M.-D. DEVIGNES. *A data science approach for exploring differential expression profiles of genes in transcriptomic studies-Application to the understanding of ageing in obese and lean rats in the FIGHT-HF project*, July 2018, JOBIM 2018 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster, https://hal.inria.fr/hal-01928421

[29] I. CHAUVOT DE BEAUCHÊNE, S. J. DE VRIES. *Reactive pipelines for integrated structural bioinformatics resources*, October 2018, 3D-BioInfo: Launch Meeting for a proposed ELIXIR Community in Structural Bioinformatics, Poster, https://hal.inria.fr/hal-01925064

[30] D. RITCHIE. *Whole Number Recursion Formulae for High Order Clebsch-Gordan Coupling Coefficients*, July 2018, working paper or preprint, https://hal.inria.fr/hal-01851097

[31] M. E. RUIZ ECHARTEA, I. CHAUVOT DE BEAUCHÊNE, D. RITCHIE. *EROS: A Protein Docking Algorithm Using a Quaternion Pi-Ball Representation for Exhaustive and Accelerated Exploration of 3D Rotational Space*, November 2018, APIL 2018 - Journée d'automne de l'Ecole Doctorale IAEM-Lorraine, Poster, https://hal.inria.fr/hal-01929546

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

[33] C. SINGHAL, Y. PONTY, I. CHAUVOT DE BEAUCHÊNE. *A hybrid combinatorial method for docking single stranded RNA on proteins at the thermodynamic equilibrium*, April 2018, RECOMB 2018 - 22nd Annual International Conference on Research in Computational Molecular Biology, Poster, https://hal.inria.fr/hal-01925083

References in notes

[34] S. Z. ALBORZI, S. ARIDHI, M.-D. DEVIGNES, R. SAIDI, A. RENAUX, M. J. MARTIN, D. W. RITCHIE. *Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations*, in "Function-SIG ISMB/ECCB 2017", Prague, Czech Republic, biofunctionprediction.org, July 2017, https://hal.inria.fr/hal-01573070

[35] S. Z. ALBORZI, M.-D. DEVIGNES, S. ARIDHI, R. SAIDI, A. RENAUX, M. J. MARTIN, D. W. RITCHIE. *Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations*, July 2017, Function-SIG ISMB/ECCB 2017, Poster, https://hal.inria.fr/hal-01573079

[36] S. Z. ALBORZI, M.-D. DEVIGNES, D. RITCHIE. *Associating Gene Ontology Terms with Pfam Protein Domains*, in "5th International Work-Conference on Bioinformatics and Biomedical Engineering - IWBBIO 2017", Granada, Spain, I. ROJAS, F. ORTUÑO (editors), Bioinformatics and Biomedical Engineering, Springer, April 2017, vol. 10209, pp. 127-138 [*DOI :* 10.1007/978-3-319-56154-7_13], https://hal.inria.fr/hal-01531204

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

[38] S. ARIDHI, A. MONTRESOR, Y. VELEGRAKIS. *BLADYG: A Graph Processing Framework for Large Dynamic Graphs*, in "Big Data Research", 2017 [*DOI :* 10.1016/J.BDR.2017.05.003], https://hal.inria.fr/hal-01577882

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

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

[41] H. M. BERMAN, T. BATTISTUZ, T. N. BHAT, W. F. BLUHM, P. E. BOURNE, K. BURKHARDT, Z. FENG, G. L. GILLILAND, L. IYPE, S. JAIN, P. FAGAN, J. MARVIN, D. PADILLA, V. RAVICHANDRAN, B.

SCHNEIDER, N. THANKI, H. WEISSIG, J. D. WESTBROOK, C. ZARDECKI. *The Protein Data Bank*, in "Acta. Cryst.", 2002, vol. D58, pp. 899–907

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

[43] E. BRESSO, V. LEROUX, M. URBAN, K. HAMMOND-KOSACK, B. MAIGRET, N. F. MARTINS. *Structure-based virtual screening of hypothetical inhibitors of the enzyme longiborneol synthase-a potential target to reduce Fusarium head blight disease*, in "Journal of Molecular Modeling", July 2016, vol. 22, n^{o} 7 [*DOI :* 10.1007/S00894-016-3021-1], https://hal.inria.fr/hal-01392851

[44] I. CHAUVOT DE BEAUCHÊNE, S. J. DE VRIES, M. ZACHARIAS. *Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins*, in "Nucleic Acids Research", June 2016 [*DOI :* 10.1093/NAR/GKW328], https://hal.archives-ouvertes.fr/hal-01505862

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

[46] C. COLUZZI, G. GUÉDON, M.-D. DEVIGNES, C. AMBROSET, V. LOUX, S. PAYOT, N. N. LEBLOND-BOURGET, T. LACROIX. *A Glimpse into the World of Integrative and Mobilizable Elements in Streptococci Reveals an Unexpected Diversity and Novel Families of Mobilization Proteins*, in "Frontiers in Microbiology", March 2017, vol. 8, 16 p. [*DOI :* 10.3389/FMICB.2017.00443], https://hal.inria.fr/hal-01580789

[47] P. COUVINEAU, H. DE ALMEIDA, B. MAIGRET, C. LLORENS-CORTES, X. ITURRIOZ. *Involvement of arginine 878 together with Ca2+ in aminopeptidase A substrate specificity for N-terminal acidic amino-acid residues*, in "PLoS One", September 2017, https://hal.inria.fr/hal-01580832

[48] K. DALLEAU, M. COUCEIRO, M.-D. DEVIGNES, C. RAÏSSI, M. SMAÏL-TABBONE. *Using aggregation functions on structured data: a use case in the FIGHT-HF project*, in "International Symposium on Aggregation and Structures (ISAS 2016) ", Luxembourg, Luxembourg, G. KISS, J.-L. MARICHAL, B. TEHEUX (editors), International Symposium on Aggregation and Structures (ISAS 2016) - Book of abstracts, July 2016, https://hal.inria.fr/hal-01399232

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

[50] M.-D. DEVIGNES, B. SIDAHMED, M. SMAÏL-TABBONE, N. AMEDEO, P. OLIVIER. *Functional classification of genes using semantic distance and fuzzy clustering approach: Evaluation with reference sets and overlap analysis*, in "international Journal of Computational Biology and Drug Design. Special Issue on: "Systems Biology Approaches in Biological and Biomedical Research"", 2012, vol. 5, n^{o} 3/4, pp. 245-260, https://hal.inria.fr/hal-00734329

[51] W. DHIFLI, S. ARIDHI, E. MEPHU NGUIFO. *MR-SimLab: Scalable subgraph selection with label similarity for big data*, in "Information Systems", 2017, vol. 69, pp. 155 - 163 [*DOI :* 10.1016/J.IS.2017.05.006], https://hal.inria.fr/hal-01573398

[52] S. E. DOBBINS, V. I. LESK, M. J. E. STERNBERG. *Insights into protein flexibility: The relationship between normal modes and conformational change upon protein–protein docking*, in "Proceedings of National Academiy of Sciences", 2008, vol. 105, n° 30, pp. 10390–10395

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

[54] R. D. FINN, J. MISTRY, J. TATE, P. COGGILL, A. HEGER, J. E. POLLINGTON, O. L. GAVIN, P. GUNASEKARAN, G. CERIC, K. FORSLUND, L. HOLM, E. L. L. SONNHAMMER, S. R. EDDY, A. BATEMAN. *The Pfam protein families database*, in "Nucleic Acids Research", 2010, vol. 38, pp. D211–D222

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

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

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

[58] H. HERMJAKOB, L. MONTECCHI-PALAZZI, G. BADER, J. WOJCIK, L. SALWINSKI, A. CEOL, S. MOORE, S. ORCHARD, U. SARKANS, C. VON MERING, B. ROECHERT, S. POUX, E. JUNG, H. MERSCH, P. KERSEY, M. LAPPE, Y. LI, R. ZENG, D. RANA, M. NIKOLSKI, H. HUSI, C. BRUN, K. SHANKER, S. G. N. GRANT, C. SANDER, P. BORK, W. ZHU, A. PANDEY, A. BRAZMA, B. JACQ, M. VIDAL, D. SHERMAN, P. LEGRAIN, G. CESARENI, I. XENARIOS, D. EISENBERG, B. STEIPE, C. HOGUE, R. APWEILER. *The HUPO PSI's Molecular Interaction format – a community standard for the representation of protein interaction data*, in "Nature Biotechnology", 2004, vol. 22, n° 2, pp. 177-183

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

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

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

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

[63] N. F. MARTINS, E. BRESSO, R. C. TOGAWA, M. URBAN, J. ANTONIW, B. MAIGRET, K. HAMMOND-KOSACK. *Searching for Novel Targets to Control Wheat Head Blight Disease-I-Protein Identification, 3D Modeling and Virtual Screening*, in "Advances in Microbiology", September 2016, vol. 06, n° 11, pp. 811 - 830 [*DOI :* 10.4236/AIM.2016.611079], https://hal.inria.fr/hal-01392860

[64] L. MAVRIDIS, V. VENKATRAMAN, D. W. RITCHIE. *A Comprehensive Comparison of Protein Structural Alignment Algorithms*, in "3DSIG – 8th Structural Bioinformatics and Computational Biophysics Meeting", Long Beach, California, ISMB, 2012, vol. 8, 89 p.

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

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

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

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

[69] S. ORCHARD, S. KERRIEN, S. ABBANI, B. ARANDA, J. BHATE, S. BIDWELL, A. BRIDGE, L. BRIGANTI, F. S. L. BRINKMAN, G. CESARENI, A. CHATRARYAMONTRI, E. CHAUTARD, C. CHEN, M. DUMOUSSEAU, J. GOLL, R. E. W. HANCOCK, L. I. HANNICK, I. JURISICA, J. KHADAKE, D. J. LYNN, U. MAHADEVAN, L. PERFETTO, A. RAGHUNATH, S. RICARD-BLUM, B. ROECHERT, L. SALWINSKI, V. STÜMPFLEN, M. TYERS, P. UETZ, I. XENARIOS, H. HERMJAKOB. *Protein interaction data curation: the International Molecular Exchange (IMEx) consortium*, in "Nature Methods", 2012, vol. 9, n^{o} 4, pp. 345-350

[70] G. PERSONENI, E. BRESSO, M.-D. DEVIGNES, M. DUMONTIER, M. SMAÏL-TABBONE, A. COULET. *Discovering associations between adverse drug events using pattern structures and ontologies*, in "Journal of Biomedical Semantics", 2017, vol. 93, pp. 539 - 546 [*DOI :* 10.1038/CLPT.2013.24], https://hal.inria.fr/hal-01576341

[71] G. PERSONENI, S. DAGET, C. BONNET, P. JONVEAUX, M.-D. DEVIGNES, M. SMAÏL-TABBONE, A. COULET. *ILP for Mining Linked Open Data: a biomedical Case Study*, in "The 24th International Conference on Inductive Logic Programming (ILP 2014)", Nancy, France, September 2014, https://hal.inria.fr/hal-01095597

[72] G. PERSONENI, S. DAGET, C. BONNET, P. JONVEAUX, M.-D. DEVIGNES, M. SMAÏL-TABBONE, A. COULET. *Mining Linked Open Data: A Case Study with Genes Responsible for Intellectual Disability*, in "Data Integration in the Life Sciences - 10th International Conference, DILS 2014", Lisbon, Portugal, H. GALHARDAS, E. RAHM (editors), Lecture Notes in Computer Science, Springer, 2014, vol. 8574, pp. 16 - 31, https://hal.inria.fr/hal-01095591

[73] G. PERSONENI, M.-D. DEVIGNES, M. DUMONTIER, M. SMAÏL-TABBONE, A. COULET. *Discovering ADE associations from EHRs using pattern structures and ontologies*, in "Phenotype Day, Bio-Ontologies SIG, ISMB", Orlando, United States, July 2016, https://hal.inria.fr/hal-01369448

[74] G. PERSONENI, M.-D. DEVIGNES, M. DUMONTIER, M. SMAÏL-TABBONE, A. COULET. *Extraction d'association d'EIM à partir de dossiers patients : expérimentation avec les structures de patrons et les ontologies*, in "Deuxième Atelier sur l'Intelligence Artificielle et la Santé", Montpellier, France, Atelier IA & Santé, June 2016, https://hal.inria.fr/hal-01391172

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

[76] D. RITCHIE, A. GHOORAH, L. MAVRIDIS, V. VENKATRAMAN. *Fast Protein Structure Alignment using Gaussian Overlap Scoring of Backbone Peptide Fragment Similarity*, in "Bioinformatics", October 2012, vol. 28, n^{o} 24, pp. 3274-3281 [*DOI :* 10.1093/BIOINFORMATICS/BTS618], https://hal.inria.fr/hal-00756813

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

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

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

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

[81] S. ROSELLI, A. OLRY, S. VAUTRIN, O. O. CORITON, D. RITCHIE, G. GALATI, N. NAVROT, C. KRIEGER, G. VIALART, H. H. BERGES, F. BOURGAUD, A. HEIN. *A bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip*, in "Plant Journal", February 2017, vol. 89, n^{o} 6, pp. 1119-1132 [*DOI :* 10.1111/TPJ.13450], https://hal.inria.fr/hal-01531248

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

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

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

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

[86] S. VELANKAR, J. M. DANA, J. JACOBSEN, G. VAN GINKEL, P. J. GANE, J. LUO, T. J. OLDFIELD, C. O'DONOVAN, M.-J. MARTIN, G. J. KLEYWEGT. *SIFTS: Structure Integration with Function, Taxonomy and Sequences resource*, in "Nucleic Acids Research", 2012, vol. 41, pp. D483–D489

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

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

[89] S. YANG, P. E. BOURNE. *The Evolutionary History of Protein Domains Viewed by Species Phylogeny*, in "PLoS One", 2009, vol. 4, e8378 p.

[90] Q. C. ZHANG, D. PETREY, L. DENG, L. QIANG, Y. SHI, C. A. THU, B. BISIKIRSKA, C. LEFEBVRE, D. ACCILI, T. HUNTER, T. MANIATIS, A. CALIFANO, B. HONIG. *Structure-based prediction of protein-protein interactions on a genome-wide scale*, in "Nature", 2012, vol. 490, pp. 556–560

[91] H. DE ALMEIDA, V. LEROUX, F. N. MOTTA, P. GRELLIER, B. MAIGRET, J. M. SANTANA, I. M. D. BASTOS. *Identification of novel Trypanosoma cruzi prolyl oligopeptidase inhibitors by structure-based virtual screening*, in "Journal of Computer-Aided Molecular Design", October 2016 [*DOI :* 10.1007/s10822-016-9985-1], https://hal.inria.fr/hal-01392842

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