IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2016

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

1. Members **1**
2. Overall Objectives **2**
3. Research Program **3**
3.1. Classifying and Mining Protein Structures and Protein Interactions 3
3.1.1. Context 3
3.1.2. Quantifying Structural Similarity 3
3.1.3. Formalising and Exploiting Domain Knowledge 3
3.1.4. 3D Protein Domain Annotation and Shape Mining 4
3.2. Integrative Multi-Component Assembly and Modeling 4
3.2.1. Context 4
3.2.2. Polar Fourier Docking Correlations 4
3.2.3. Assembling Symmetrical Protein Complexes 5
3.2.4. Coarse-Grained Models 5
3.2.5. Assembling Multi-Component Complexes and Integrative Structure Modeling 6
4. Application Domains **6**
4.1. Biomedical Knowledge Discovery 6
4.2. Prokaryotic Type IV Secretion Systems 7
4.3. G-protein Coupled Receptors 7
5. Highlights of the Year **7**
6. New Software and Platforms **8**
6.1. Hex 8
6.2. Kbdock 8
6.3. Kpax 8
6.4. Sam 8
6.5. ECDomainMiner 8
6.6. Platforms 9
7. New Results **9**
7.1. Correlating Adverse Drug Side Effects 9
7.2. Docking Symmetrical Protein Structures 9
7.3. Multiple Flexible Protein Structure Alignments 9
7.4. Annotating 3D Protein Domains 10
7.5. Identifying New Anti-Fungal Agents 10
8. Partnerships and Cooperations **10**
8.1. Regional Initiatives 10
8.2. National Initiatives 10
8.2.1. FEDER 10
8.2.2. ANR 10
8.2.2.1. Fight-HF 10
8.2.2.2. IFB 10
8.2.2.3. PEPSI 11
8.3. International Initiatives 11
8.4. International Research Visitors 11
9. Dissemination **11**
9.1. Promoting Scientific Activities 11
9.1.1. Scientific Events Organisation 11
9.1.1.1. General Chair, Scientific Chair 11
9.1.1.2. Member of Organizing Committees 12
9.1.2. Scientific Events Selection 12
9.1.2.1. Member of Conference Program Committees 12

9.1.2.2. Reviewer 12
9.1.3. Journal 12
9.1.3.1. Member of Editorial Boards 12
9.1.3.2. Reviewing Activities 12
9.1.4. Invited Talks 12
9.1.5. Research Administration 12
9.2. Teaching - Supervision - Juries 12
9.2.1. Teaching 12
9.2.2. Supervision 12
9.2.3. Juries 13
9.3. Popularization 13
10. Bibliography **13**

Project-Team CAPSID

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

Keywords:

Computer Science and Digital Science:

1.5.1. - Systems of systems
3.1.1. - Modeling, representation
3.2.2. - Knowledge extraction, cleaning
3.2.5. - Ontologies
6.1.5. - Multiphysics modeling

Other Research Topics and Application Domains:

1.1.1. - Structural biology
1.1.2. - Molecular biology
1.1.9. - Bioinformatics
2.2.1. - Cardiovascular and respiratory diseases
2.2.4. - Infectious diseases, Virology

1. Members

Research Scientists

David Ritchie [Team leader, Inria, Senior Researcher, HDR]
Isaure Chauvot de Beauchêne [CNRS, Researcher, from Dec 2016]
Marie-Dominique Devignes [CNRS, Researcher, HDR]
Bernard Maigret [CNRS, Senior Researcher, HDR]

Faculty Member

Sabeur Aridhi [Univ. Lorraine, Associate Professor, from Sep 2016]

Engineer

Antoine Chemardin [Inria]

PhD Students

Seyed Ziaeddin Alborzi [Inria]
Gabin Personeni [Univ. Lorraine]
Maria Elisa Ruiz Echartea [Inria, from Nov 2016]

Administrative Assistants

Emmanuelle Deschamps [Inria]
Laurence Félicité [Univ. Lorraine]
Christelle Levêque [Univ. Lorraine]

Others

Jessica Cummins [Univ. Dublin, Student, from Apr 2016 until Aug 2016]
Pierre Daudier de Cassini [INSERM, Student, from Apr 2016 until Jun 2016]
Marwa El Houasli [Univ. Paris 7, Student, from Apr 2016 until Aug 2016]
Valerian Gonnot [CPP-INP, Student, from May 2016 until Jun 2016]
Maxime Guyot [INSERM, Student, from May 2016 until Aug 2016]
Vincent Leroux [Univ. Lorraine, Visiting Scientist, from Apr 2016 until Dec 2016]
Zacharie Mehnana [CNRS, Student, from Jun 2016 until Aug 2016]

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 and RNA may be considered to encode the blueprint for life, whereas proteins make up the three-dimensional (3D) molecular machinery. Many biological processes are governed by complex systems of proteins which interact cooperatively to regulate the chemical composition within a cell or to carry out a wide range of biochemical processes such as photosynthesis, metabolism, and cell signalling, for example. It is becoming increasingly feasible to isolate and characterise some of the individual protein components of such systems, but it still remains extremely difficult to achieve detailed models of how these complex systems actually work. Consequently, a new multidisciplinary approach called integrative structural biology has emerged which aims to bring together experimental data from a wide range of sources and resolution scales in order to meet this challenge [69], [53].

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 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 [66], [44]. 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 [48], [71]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [34].

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 [70]. 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 [59]. 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 [50]. The latest version of Kpax [20] 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 [36], 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) [40].

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 [45] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [55]. A different point of view is represented in the Interaction Network Ontology (INO; http://www.ino-ontology.org/) which is a community-driven ontology that is being developed to standardise and integrate data on interaction networks and to support computer-assisted reasoning [72]. 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" [39], 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 [1], [3], 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) [67] functional annotations between standard Uniprot (http://www.uniprot.org/ sequence identifiers and protein structures from the Protein Data Bank (PDB) [33] 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 have developed a recommender-based data mining technique to associate enzyme classification code numbers with Pfam domains using our recently developed EC-DomainMiner program [29].

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 [60], the shape of a protein molecule is represented using polar Fourier series expansions of the form

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

where $\sigma(\underline{x})$ is a 3D shape-density function, a_{nlm} are the expansion coefficients, $R_{nl}(r)$ are orthonormal Gauss-Laguerre polynomials and $y_{lm}(\theta,\phi)$ are the real spherical harmonics. The electrostatic potential, $\phi(\underline{x})$, and charge density, $\rho(\underline{x})$, of a protein may be represented using similar expansions. Such representations allow the *in vacuo* electrostatic interaction energy between two proteins, A and B, to be calculated as [47]

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

This equation demonstrates using the notion of *overlap* between 3D scalar quantities to give a physics-based scoring function. If the aim is to find the configuration that gives the most favourable interaction energy, then it is necessary to perform a six-dimensional search in the space of available rotational and translational degrees of freedom. By re-writing the polar Fourier expansions using complex spherical harmonics, we showed previously that fast Fourier transform (FFT) techniques may be used to accelerate the search in up to five of the six degrees of freedom [61]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [9], [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 [62], [49], 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 dramatically [61][9]. 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 [32], [58], [65]. 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

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

where the identical monomers A and B are initially placed at the origin, and $\omega_n = 2\pi/n$ is the rotation about the principal n-fold symmetry axis. This example shows that complexes with cyclic symmetry have just 4 rigid body degrees of freedom (DOFs), compared to $6(n-1)$ DOFs for non-symmetrical n-mers. We have generalised these ideas in order to model protein complexes that crystallise into any of the naturally occurring point group symmetries (C_n, D_n, T, O, I). This approach was published in 2016 [19], and was subsequently applied to several symmetrical complexes from the "CAPRI" blind docking experiment [13]. Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to 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 [54], [37], [51], [52]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [68], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [2].

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 [31]. 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 [64]. Furthermore, this kind of coarse-graining effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [46]. 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 (commenced November 2016).

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, its 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. Although we do not have precise road-map to a solution for the multi-component assembly problem, we wish to proceed firstly by putting more emphasis on the single-body terms in the scoring function, and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space.

4. Application Domains

4.1. Biomedical Knowledge Discovery

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

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 [56], [57]. 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).

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 [25]. 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.

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 [30]. 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 [41], [63]. 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 "MBI" platform (see Section 6.8), 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.

4.3. G-protein Coupled Receptors

Participants: Bernard Maigret [contact person], David Ritchie, Vincent Leroux.

G-protein coupled receptors (GPCRs) are cell surface proteins which detect chemical signals outside a cell and which transform these signals into a cascade of cellular changes. Historically, the most well documented signaling cascade is the one driven by G-proteins trimers (guanine nucleotide binding proteins) [43] which ultimately regulate many cellular processes such as transcription, enzyme activity, and homeostatis, for example. But other pathways have recently been associated with the signals triggered by GPCRs, involving other proteins such as arrestins and kinases which drive other important cellular activities. For example, β-arrestin activation can block GPCR-mediated apoptosis (cell death). Malfunctions in such processes are related to diseases such as diabetes, neurological disorders, cardiovascular disease, and cancer. Thus, GPCRs are one of the main protein families targeted by therapeutic drugs [38] and the focus of much bio-medical research. Indeed, approximately 40–50% of current therapeutic molecules target GPCRs. However, despite enormous efforts, the main difficulty here is the lack of experimentally solved 3D structures for most GPCRs. Hence, computational modeling tools are widely recognized as necessary to help understand GPCR functioning and thus biomedical innovation and drug design.

In collaboration with the BIOS team (INRA Tours) and the AMIB team (Inria Saclay – Île de France) we used our Hex protein docking software to help model a multi-component G-protein coupled receptor (GPCR) complex [35]. The resulting 3D structure was shown to be consistent with the known experimental data for the protein components of this trans-membrane molecular signaling system. As part of an on-going collaboration with the Centre for Interdisciplinary Research (CIRB) at Collège de France, we modeled the interaction between the Apelin peptide and a GPCR called ApelinR [42]. This study provided mechanistic insights which could lead to the development of therapeutic agents for the treatment of heart failure.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

A figure from our article in the *Journal of Applied Crystallography* [19] was used to illustrate the front cover of the February issue of the journal.

6. New Software and Platforms

6.1. Hex

KEYWORDS: 3D rendering - Bioinformatics - 3D interaction - Structural Biology
SCIENTIFIC DESCRIPTION The underlying algorithm uses a novel polar Fourier correlation technique to accelerate the search for close-fitting orientations of the two molecules.
FUNCTIONAL DESCRIPTION Hex is an interactive protein docking and molecular superposition program. Hex understands protein and DNA structures in PDB format, and it can also read small-molecule "SDF" files. Hex will run on most Windows-XP, Linux and Mac OS X PCs. 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). On multi-processor Linux systems, docking calculation times can be reduced in almost direct proportion to the number of CPUs and GPUs used. The calculations can be accelerated by using an optional disc cache (strongly recommended) of about 1 GB of disc space.

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

6.2. Kbdock

FUNCTIONAL DESCRIPTION Database 3D protein domain-domain interactions with a web interface

- Authors: Anisah Ghoorah, 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 To align and superpose the 3D structures of protein molecules.

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

6.4. Sam

Symmetry Assembler
FUNCTIONAL DESCRIPTION To predict the three-dimentional structures of symmetrical protein complexes using spherical polar Fourier representations

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

6.5. ECDomainMiner

KEYWORDS: Protein Domain Annotation

SCIENTIFIC DESCRIPTION

EC-DomainMiner is a recommender-based approach for associating EC (Enzyme Commission) numbers with Pfam domains.

FUNCTIONAL DESCRIPTION

EC-DomainMiner uses a statistical recommender-based approach to infer EC-Pfam relationships from EC-sequence relationships that have been annotated previously in the SIFTS and Uniprot databases.

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

6.6. Platforms

6.6.1. 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).

- Contact: Marie-Dominique Devignes

7. New Results

7.1. Correlating Adverse Drug Side Effects

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 Prof. Michel Dumontier of the 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 [26], [27]. A paper describing this work has been submitted to the Journal of Biomedical Semantics.

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 [19]. An article describing the results obtained when using Sam to dock several symmetrical protein complexes from the "CAPRI" docking experiment has also been published [13].

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. We have recently extended our "Kpax" protein structure alignment algorithm to flexibly align pairs of structures that cannot be completely superposed by a single rigid-body transformation, and to calculate multiple alignments of several similar structures flexibly. A journal article describing the approach has been published [20].

7.4. Annotating 3D Protein Domains

Many protein chains in the Protein Data Bank (PDB) are cross-referenced with EC numbers and Pfam domains. However, these annotations do not explicitly indicate any relation between EC numbers and Pfam domains. In order to address this limitation, we developed EC-DomainMiner, a recommender-based approach for associating EC (Enzyme Commission) numbers with Pfam domains [29]. EC-DomainMiner is able to infer automatically 20,179 associations between EC numbers and Pfam domains from existing EC-chain/Pfam-chain associations from the SIFTS database as well as EC-sequence/Pfam-sequence associations from UniProt databases. A manuscript describing this work has been provisionally accepted by the journal *BMC-Bioinformatics.*

7.5. Identifying New Anti-Fungal Agents

In this collaboration with several Brasilian laboratories (at University of Mato Grosso State, University of Maringá, Embrapa, and University of Brasilia), we identified several novel small-molecule drug leads against *Trypanosoma cruzi*, a parasite responsible for Chagas disease [21]. We also proposed several small-molecule inhibitors against *Fusarium graminearum*, a fungal threat to global wheat production [15], [12].

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. PEPS

Participants: Marie-Dominique Devignes [contact person], Bernard Maigret, David Ritchie.

The team is involved in the inter-disciplinary "MODEL-ICE" project led by Nicolas Soler (DynAMic lab, UMR 1128, INRA / Univ. Lorraine). 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.

8.2. National Initiatives

8.2.1. FEDER

8.2.1.1. LBS

Participant: Marie-Dominique Devignes [contact person].

The project "LBS" (Le Bois Santé) is a consortium funded by the European Regional Development Fund (FEDER) and the French "Fonds Unique Interministériel" (FUI). The project is coordinated by Harmonic Pharma SAS. The aim of LBS is to exploit wood products in the pharmaceutical and nutrition domains. Our contribution has been in data management and knowledge discovery for new therapeutic applications.

8.2.2. ANR

8.2.2.1. Fight-HF

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

This "Investissements d'Avenirs" 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.

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.

8.2.2.3. PEPSI

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

The PEPSI ("Polynomial Expansions of Protein Structures and Interactions") project is a collaboration with Sergei Grudinin at Inria Grenoble – Rône Alpes (project Nano-D) and Valentin Gordeliy at the Institut de Biologie Structurale (IBS) in Grenoble. This project funded by the ANR "Modèles Numériques" program involves developing computational protein modeling and docking techniques and using them to help solve the structures of large molecular systems experimentally.

8.3. International Initiatives

8.3.1. Participation in Other International Programs

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, Brasil; Funding: CNPq.

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

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

Participant: Bernard Maigret; Project: *Protein-protein interactions for the development of new drugs;* Partner: Federal University of Goias, Brasil. Funding: Chamada MCTI/CNPq/FNDCT.

8.4. International Research Visitors

8.4.1. Visits to International Teams

8.4.1.1. Research Stays Abroad

Gabin Personeni visited the Biomedical Informatics Research Laboratory of Prof. Michel Dumontier at Stanford University for 3 months (Nov 2015 – Feb 2016).
Seyed Ziaeddin Alborzi visited the UniProt development team of Maria Martin at the European Bioinformatics Institute (EBI), Hinxton UK, for 3 months (Oct – Dec 2016) in partial fulfilment of the requirements for a European PhD.

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).

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

Marie-Dominique Devignes organised a workshop ("Atelier Santé") for the Fédération Charles Hermite ("Journée Enterprises", 21/01/2016).

9.1.1.2. Member of Organizing Committees

Marie-Dominique Devignes co-organised a workshop on Structural Modeling of Type IV Secretion Systems (PEPS workshop "Model-ICE", 13/12/2016).

9.1.2. Scientific Events Selection

9.1.2.1. Member of Conference Program Committees

Marie-Dominique Devignes was a member of the programme committee for KDIR-2016, ECCB-2016, and BIBM-2016.

9.1.2.2. Reviewer

David Ritchie was a reviewer for IJCAI-2016.

Marie-Dominique Devignes reviewed grant applications for the Medical Research Council (UK) and National Science Centre (Poland).

9.1.3. Journal

9.1.3.1. Member of Editorial Boards

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

9.1.3.2. Reviewing Activities

The members of the team have reviewed manuscripts for *Algorithms for Molecular Biology, Bioinformatics, Current Opinion in Structural Biology, Journal of Biomedical Semantics, Journal of Computational Chemistry Journal of Chemical Information and Modeling, Journal of Molecular Recognition,* and *Proteins: Structure, Function & Bioinformatics.*

9.1.4. Invited Talks

David Ritchie gave a presentation at the *6th CAPRI Evaluation Meeting* in Tel Aviv.

9.1.5. 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, and is a member of the Bureau of the Project Committee for Inria Nancy – Grand Est.
Marie-Dominique Devignes was a member of the "Commission de spécialistes" for the recruitment of an associate professor in computer science at Telecom Nancy, Université de Lorraine, April-May 2016.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence: Sabeur Aridhi, *Programming Techniques and Tools,* 24 hours, L1, Telecom Nancy, Univ. Lorraine.

Licence: Sabeur Aridhi, *Big Data Hackathon,* 4 hours, L3, Telecom Nancy, Univ. Lorraine.

Licence: Marie-Dominique Devignes, *Relational Database Design and SQL,* 30 hours, L3, Telecom Nancy, Univ. Lorraine.

Master: Marie-Dominique Devignes, *Gene Discovery in Biological Databases,* 8 hours, M1, Univ. Lorraine.

9.2.2. Supervision

PhD in progress: Maria Elisa Ruiz Echartea, *Multi-component protein assembly using distance constraints,* 01/11/2016, David Ritchie.

PhD in progress: Gabin Personeni, *Exploration of linked open data in view of knowledge discovery. Application to the biomedical domain,* 01/10/2014, Marie-Dominique Devignes, Adrien Coulet.

PhD in progress: Seyed Ziaeddin Alborzi, *Large-scale exploration of 3D protein domain family binding sites,* 01/10/2014, David Ritchie, Marie-Dominique Devignes.

9.2.3. Juries

HDR: Olivier Dameron, *Ontology-based methods for analysing life science data,* Université de Rennes, 11/01/2016.

PhD: Minh-Son Phan, *Contribution á l'estimation de la similarité dans un ensemble de projections tomographiques non-orientées,* Université de Strasbourg, 07/10/2016, Pr Mohamed Tajine, Dr Étienne Baudrier, Dr Loïc Mazo.

PhD: Yassine Ghouzam, *Nouvelles approches pour l'analyse et la prédiction de la structure tridimensionnelle des protéines,* Université Paris 7, 18/10/2016, Dr Jean-Christophe Gelly.

PhD: Benoît Henry, *Probability theory applied to evolutionary biology,* Université de Lorraine, 17/11/2016, Dr Nicolas Champagnat, Dr David Ritchie.

PhD: Yoann Dufresne, *Algorithmique pour l'annotation automatique de peptides non ribosomiques,* Université de Lille, 01/12/2016, Pr Maude Pupin, Dr Laurent Noé.

9.3. Popularization

An article on our KBDOCK resource for studying protein-protein interactions was published in ERCIM News (edition 104, January 2016) [22].

The team made a presentation at the public "Cité Forum" in Nancy (01-02 Apr 2016).

10. Bibliography

Major publications by the team in recent years

[1] 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^{o} 20, pp. 2820-2827 [*DOI :* 10.1093/BIOINFORMATICS/BTR493], https://hal.inria.fr/inria-00617921

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

[3] 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^{o} D1, pp. 389-395, https://hal.inria.fr/hal-00920612

[4] 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

[5] T. HOANG, X. CAVIN, P. SCHULTZ, D. RITCHIE. *gEMpicker: a highly parallel GPU-accelerated particle picking tool for cryo-electron microscopy*, in "BMC Structural Biology", 2013, vol. 13, n^{o} 1, 25 p. [*DOI :* 10.1186/1472-6807-13-25], https://hal.inria.fr/hal-00955580

[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. 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

[9] 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

[10] V. VENKATRAMAN, D. W. RITCHIE. *Predicting Multi-component Protein Assemblies Using an Ant Colony Approach*, in "International Journal of Swarm Intelligence Research", September 2012, vol. 3, pp. 19-31 [*DOI :* 10.4018/JSIR.2012070102], https://hal.inria.fr/hal-00756807

Publications of the year

Articles in International Peer-Reviewed Journals

[11] C. AMBROSET, C. COLUZZI, G. GUÉDON, M.-D. DEVIGNES, V. LOUX, T. LACROIX, S. PAYOT, N. LEBLOND-BOURGET. *New Insights into the Classification and Integration Specificity of Streptococcus Integrative Conjugative Elements through Extensive Genome Exploration*, in "Frontiers in microbiology", January 2016, vol. 6, 1483 p. [*DOI :* 10.3389/FMICB.2015.01483], https://hal.archives-ouvertes.fr/hal-01262284

[12] E. BRESSO, V. LEROUX, M. URBAN, K. E. 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° 7 [*DOI :* 10.1007/S00894-016-3021-1], https://hal.inria.fr/hal-01392851

[13] 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", October 2016 [*DOI :* 10.1002/PROT.25182], https://hal.inria.fr/hal-01388654

[14] M. F. LENSINK, S. VELANKAR, A. KRYSHTAFOVYCH, S.-Y. HUANG, D. SCHNEIDMAN-DUHOVY, A. SALI, J. SEGURA, N. FERNANDEZ-FUENTES, S. VISWANATH, R. ELBER, S. GRUDININ, P. POPOV, E. NEVEU, H. LEE, M. BAEK, S. PARK, L. HEO, G. R. LEE, C. SEOK, S. QIN, H.-X. ZHOU, D. W. RITCHIE, B. MAIGRET, M.-D. DEVIGNES, A. GHOORAH, M. TORCHALA, R. A.G. CHALEIL, P. A. BATES, E. BEN-ZEEV, M. EISENSTEIN, S. NEGI S., T. VREVEN, B. G. PIERCE, T. M. BORRMAN, J. YU, F. OCHSENBEIN, Z. WENG, R. GUEROIS, A. VANGONE, J. P. RODRIGUES, G. VAN ZUNDERT, M. NELLEN, L. XUE, E. KARACA, A. S. J. MELQUIOND, K. VISSCHER, P. L. KASTRITIS, A. M. J. J. BONVIN, X. XU, L. QIU, C. YAN, J. LI, Z. MA, J. CHENG, X. ZOU, Y. SHENG, L. X. PETERSON, H.-R. KIM, A. ROY, X. HAN, J. ESQUIVEL-RODRÍGUEZ, D. KIHARA, X. YU, N. J. BRUCE, J. C. FULLER, R. C. WADE, I. ANISHCHENKO, P. J. KUNDROTAS, I. A. VAKSER, K. IMAI, K. YAMADA, T. ODA, T. NAKAMURA,

K. TOMII, C. PALLARA, M. ROMERO-DURANA, B. JIMÉNEZ-GARCÍA, I. H. MOAL, J. FERNÁNDEZ-RECIO, J. Y. JOUNG, J. Y. KIM, K. JOO, J. LEE, D. KOZAKOV, S. VAJDA, S. MOTTARELLA, D. R. HALL, D. BEGLOV, A. MAMONOV, B. XIA, T. BOHNUUD, C. A. DEL CARPIO, E. ICHIISHI, N. MARZE, D. KURODA, S. S. R. BURMAN, J. J. GRAY, E. CHERMAK, L. CAVALLO, R. OLIVA, A. TOVCHIGRECHKO, S. J. WODAK. *Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment*, in "Proteins - Structure, Function and Bioinformatics", February 2016 [*DOI :* 10.1002/PROT.25007], https://hal.inria.fr/hal-01309105

[15] M. 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^{o} 11, pp. 811 - 830 [*DOI :* 10.4236/AIM.2016.611079], https://hal.inria.fr/hal-01392860

[16] E. NEVEU, D. RITCHIE, P. POPOV, S. GRUDININ. *PEPSI-Dock: a detailed data-driven protein–protein interaction potential accelerated by polar Fourier correlation*, in "Bioinformatics", September 2016 [*DOI :* 10.1093/BIOINFORMATICS/BTW443], https://hal.archives-ouvertes.fr/hal-01358645

[17] D. PADHORNY, A. KAZENNOV, B. S. ZERBE, K. A. PORTER, B. XIA, S. MOTTARELLA, Y. KHOLODOV, D. RITCHIE, S. VAJDA, D. KOZAKOV. *Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds*, in "PNAS Early Edition", May 2016, vol. 113, n^{o} 30, pp. E4286-E4293 [*DOI :* 10.1073/PNAS.1603929113], https://hal.inria.fr/hal-01371087

[18] M. RICHARD, A. CHATEAU, C. JELSCH, C. DIDIERJEAN, X. MANIVAL, C. CHARRON, B. MAIGRET, M. BARBERI-HEYOB, Y. CHAPLEUR, C. BOURA, N. PELLEGRINI-MOÏSE. *Carbohydrate-based peptidomimetics targeting neuropilin-1: synthesis, molecular docking study and in vitro biological activities*, in "Bioorganic and Medicinal Chemistry", November 2016, vol. 24, n^{o} 21, pp. 5315-5325 [*DOI :* 10.1016/J.BMC.2016.08.052], https://hal.archives-ouvertes.fr/hal-01360034

[19] 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^{o} 1, pp. 158-167 [*DOI :* 10.1107/S1600576715022931], https://hal.inria.fr/hal-01261402

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

[21] 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

Articles in Non Peer-Reviewed Journals

[22] M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Kbdock - Searching and organising the structural space of protein-protein interactions*, in "ERCIM News", January 2016, n^{o} 104, pp. 24-25, https://hal.inria.fr/hal-01258117

International Conferences with Proceedings

[23] C. CHAMARD-JOVENIN, A. CHESNEL, E. BRESSO, C. MOREL, C. THIÉBAUT, M. SMAÏL-TABBONE, E.-H. DJERMOUNE, M.-D. DEVIGNES, T. BOUKHOBZA, H. DUMOND. *Transgenerational effects of ER-alpha36 over-expression on mammary gland development and molecular phenotype: clinical perspective for breast cancer risk and therapy*, in "21st World Congress on Advances in Oncology and 19th International Symposium on Molecular Medicine", Athens, Greece, October 2016, https://hal.archives-ouvertes.fr/hal-01416469

[24] C. CHAMARD-JOVENIN, A. CHESNEL, C. MOREL, M.-D. DEVIGNES, M. SMAÏL-TABBONE, T. BOUKHOBZA, H. DUMOND. *Long chain alkylphenol mixture promotes breast cancer initiation and progression through an ERα36-mediated mechanism*, in "2nd French Workshop on Endocrine disruption in wildlife and human health", Paris, France, January 2016, Présentation Poster, https://hal.archives-ouvertes.fr/hal-01320688

[25] 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

National Conferences with Proceedings

[26] 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

Conferences without Proceedings

[27] 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

Scientific Books (or Scientific Book chapters)

[28] A. W. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Classification and Exploration of 3D Protein Domain Interactions Using Kbdock*, in "Data Mining Techniques for the Life Sciences", O. CARUGO, F. EISENHABER (editors), Methods in Molecular Biology, Springer Science+Business Media New York, 2016, vol. 1415, pp. 91-105 [*DOI :* 10.1007/978-1-4939-3572-7_5], https://hal.inria.fr/hal-01317448

References in notes

[29] S. Z. ALBORZI, M.-D. DEVIGNES, D. RITCHIE. *EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains*, in "JOBIM 2015", Clermont-Ferrand, France, July 2015 [*DOI :* 10.1101/022343], https://hal.inria.fr/hal-01216743

[30] 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

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

[32] 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

[33] 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

[34] 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

[35] T. BOURQUARD, F. LANDOMIEL, E. REITER, P. CRÉPIEUX, D. W. RITCHIE, J. AZÉ, A. POUPON. *Unraveling the molecular architecture of a G protein-coupled receptor/β-arrestin/Erk module complex*, in "Scientific Reports", June 2015, 5:10760 [*DOI :* 10.1038/SREP10760], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01162594

[36] 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° 3/4, pp. 245-260, https://hal.inria.fr/hal-00734329

[37] 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

[38] D. FILMORE. *It's a GPCR world*, in "Modern Drug Discovery", 2004, vol. 7, pp. 24–28

[39] 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

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

[41] 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

[42] R. GERBIER, V. LEROUX, P. COUVINEAU, R. ALVEAR-PEREZ, B. MAIGRET, C. LLORENS-CORTES, X. ITURRIOZ. *New structural insights into the apelin receptor: identification of key residues for apelin binding*, in "FASEB Journal", January 2015, vol. 29, n° 1, pp. 314-322 [*DOI :* 10.1096/FJ.14-256339], https://hal.inria.fr/hal-01251633

[43] A. G. GILMAN. *G proteins: transducers of receptor-generated signaling*, in "Annual Review of Biochemistry", 1987, vol. 56, pp. 615–649

[44] 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

[45] 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

[46] 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

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

[48] 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

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

[50] 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.

[51] 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

[52] 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° 10, pp. 3623–3648

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

[54] 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

[55] 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° 4, pp. 345-350

[56] 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

[57] 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

[58] 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

[59] 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

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

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

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

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

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

[65] 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

[66] 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

[67] 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

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

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

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

[71] 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

[72] 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.