IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2015

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:

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

Other Research Topics and Application Domains:

1. - Life sciences
1.1. - Biology
1.1.1. - Structural biology
1.1.8. - Evolutionnary biology
1.1.9. - Bioinformatics

1. Members

Research Scientists

David Ritchie [Team leader, Inria, Senior Researcher, HdR]
Marie-Dominique Devignes [CNRS, Researcher, HdR]
Bernard Maigret [CNRS, Senior Researcher, HdR]

Engineer

Jérémie Bourseau [CNRS, until Apr 2015]

PhD Students

Seyed Ziaeddin Alborzi [Inria]
Gabin Personeni [Univ. Lorraine]

Visiting Scientists

Ghania Khensous [Univ. Algérie, until Apr 2015]
Vincent Leroux [Collège de France, Jun 2015]
Ana Carolline Toledo [Univ. Brasilia, from Jun until Aug 2015]

Administrative Assistants

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

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 [57], [42].

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 [54], [32]. 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 [37], [59]. Therefore, developing techniques which can mine knowledge of protein structures and their interactions is an important way to enhance our knowledge of biology [24].

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 [58]. 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 [7]. 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 [39]. The latest version of Kpax (manuscript in review) 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 [25], 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) [29].

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 [33] developed by the Human Genome Organization (HUGO) mostly address the experimental wet lab aspects leading to data production and curation [44]. 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 [60]. 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” [28], 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) [56] functional annotations between standard Uniprot (http://www.uniprot.org/ sequence identifiers and protein structures from the Protein Databank (PDB) [23] 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 made good progress in this area by developing a recommender-based data mining technique to associate enzyme classification code numbers with Pfam domains using our recently developed EC-DomainMiner program [19].

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 [48], 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 [35]

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

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 [49]. Furthermore, we also showed that such calculations may be accelerated dramatically on modern graphics processor units [8], [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 [50], [38], 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 [49][8]. 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 [22], [47], [53]. 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

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

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 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). Although we currently use shape-based FFT correlations, the symmetry operator technique may equally be used to refine candidate solutions using a more accurate CG force-field scoring function.

3.2.4. Coarse-Grained Models

Many approaches have been proposed in the literature to take into account protein flexibility during docking. The most thorough methods rely on expensive atomistic simulations using MD. However, much of a MD trajectory is unlikely to be relevant to a docking encounter unless it is constrained to explore a putative protein-protein interface. Consequently, MD is normally only used to refine a small number of candidate rigid body docking poses. A much faster, but more approximate method is to use coarse-grained (CG) normal mode analysis (NMA) techniques to reduce the number of flexible degrees of freedom to just one or a handful of the most significant vibrational modes [43], [26], [40], [41]. In our experience, docking ensembles of NMA conformations does not give much improvement over basic FFT-based soft docking [9], 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 [21]. 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 [52]. Furthermore, this kind of coarse-graining effectively integrates out many of the internal DOFs to leave a smoother but still physically realistic energy surface [34]. We are therefore developing a “coarse-grained” scoring function for fast protein-protein docking and multi-component assembly.

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 roadmap 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], David Ritchie.

This project is in collaboration with the Orpailleur Team.

Increasing amounts of biomedical data provided as Linked Open Data (LOD) offer novel opportunities for knowledge discovery in biomedicine. We published an approach for selecting, integrating, and mining LOD with the goal of discovering genes responsible for a disease [46]. The selection step relies on a set of choices made by a domain expert to isolate relevant pieces of LOD. Because these pieces are potentially not linked, an integration step is required to connect unlinked pieces. The resulting graph is subsequently mined using Inductive Logic Programming (ILP) that presents two main advantages. First, the input format compliant with ILP (first order logic) is close to the format of LOD (RDF triples). Second, domain knowledge can be added to this input and used during the induction step. We have applied this approach to the characterization of genes responsible for intellectual disability. For this real-world use case, we could evaluate ILP results and assess the contribution of domain knowledge. Our ongoing efforts explore how the combination of rules coming from distinct theories can improve the prediction accuracy [45], [55].

4.2. Prokaryotic Type IV Secretion Systems

Participants: Marie-Dominique Devignes [contact person], Bernard Maigret, 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 [20]. 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 [30], [51]. 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 has initiated a 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 40 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, Ana Carolline Toledo.

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) [31] 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 [27] 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.

5. Highlights of the Year

5.1. Highlights of the Year

Large ANR Grant – Investissements d'Avenirs

Marie-Dominique Devignes and Malika Smaïl-Tabbone (Orpailleur Team) coordinated a work-package on network-based science for the project “FIGHT_HF” (Fight Heart Failure) that was submitted by Nancy University Hospital's Federation “CARTAGE” (http://www.fhu-cartage.com/) to the ANR “Investissements d'Avenirs” programme. This project aims to discover novel mechanisms for heart failure and to propose decision support for precision medicine. The project has been granted € 9M.

Journal Front Cover

A figure from our article in the *Journal of Chemical Information and Modeling* [15] was used to illustrate the front cover of the August issue of the journal.

6. New Software and Platforms

6.1. Kpax

KEYWORDS: Protein Structure Alignment

SCIENTIFIC DESCRIPTION

Kpax is a program for flexibly aligning two or more protein structures and for searching databases of protein structures.

FUNCTIONAL DESCRIPTION

The Kpax program exploits the fact that each amino acid residue has a carbon atom with a highly predictable tetrahedral geometry. This allows the local environment of each residue to be transformed into a canonical orientation, thus allowing easy comparison between the canonical orientations of residues within pairs of proteins using a novel scoring function based on Gaussian overlaps. Kpax is now used by the KBDOCK web server [3] to find structural templates for docking which might be beyond the reach of sequence-based homology modeling approaches. In 2015, the Kpax program was extended to allow the flexible alignment and superposition of multiple protein structures, and a new multiple alignment quality measure has been developed. According to this quality measure, Kpax gives higher quality multiple structural alignments than all other published approaches. A journal article describing these new developments is under review.

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

6.2. KBDOCK

KEYWORDS: Protein Binding Sites

SCIENTIFIC DESCRIPTION

KBDOCK is a database of all known protein-protein interactions that have experimentally determined 3D structures. In 2015, we used the latest version of KBDOCK in several rounds of the community-wide “CAPRI” blind docking experiment [36]. A journal article has been accepted for publication in *Proteins*.

FUNCTIONAL DESCRIPTION

KBDOCK combines coordinate data from the PDB with the Pfam protein domain family classification [28] in order to describe and analyze all known protein-protein interactions for which the 3D structures are available.

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

6.3. Hex

KEYWORDS: Protein Docking - 3D rendering - 3D interaction

SCIENTIFIC DESCRIPTION

Hex is an interactive protein docking and molecular superposition program. The underlying approach uses our polar Fourier correlation technique to accelerate the search for close-fitting orientations of the two protein molecules.

FUNCTIONAL DESCRIPTION

Hex understands protein and DNA structures in PDB format, and it can also read small-molecule SDF files. Hex will run on most Windows, Linux and Mac OS X computers. The recent versions 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. In 2015, the Hex code base was re-organised to separate the GUI and computational components into separate libraries. The computational library is now used in our Sam and Kpax software.

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

6.4. Sam

KEYWORDS: Protein Symmetry Assembly - Protein Docking

SCIENTIFIC DESCRIPTION

Sam is a program for building models of protein complexes having arbitrary point group symmetry. The Sam program was developed in the frame of the ANR "PEPSI" project with The Nano-D team at Inria Grenoble – Rône Alpes. A journal article describing Sam has been accepted for publication in the Journal of Applied Crystallography [16].

FUNCTIONAL DESCRIPTION

The underlying approach makes use of multiple one-dimensional polar Fourier correlations (implemented in the Hex code-base) to search rapidly a symmetry-constrained rigid body protein docking search space. The approach may be used to build symmetrical multi-component protein complexes having a given cyclic (C_n), dihedral (D_n), tetrahedral (T), octahedral (O) or icosahedral (I) point group symmetry.

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

6.5. EC-DomainMiner

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. MD-Kmean

KEYWORDS: Molecular Dynamics Analysis

SCIENTIFIC DESCRIPTION

MD-Kmean is a fast program for the analysis of large numbers of Molecular Dynamics frames. The accurate comparison of different protein structures plays important roles in structural biology, structure prediction and functional annotation. The root-mean-square-deviation (RMSD) after optimal superposition is the predominant measure of similarity due to the ease and speed of computation. MD-Kmean was designed to perform both the RMSD and the clustering step necessary to compare large numbers of protein 3D structures stored in large datasets and was applied to a set of 2 microsecond MD simulations producing 2 million frames to be compared and clustered.

FUNCTIONAL DESCRIPTION

We have implemented a very fast version of RMSD for graphics processing units (GPUs) using a quaternion method for calculating the optimal superposition and RMSD that is designed for parallel applications. This acceleration in speed allows RMSD calculations to be used efficiently in computationally intensive applications such as the clustering of large number of molecular dynamics frames. MD-Kmean is 50 times faster on a Nvidia GPU, on average, than the original single-threaded CPU implementation on an Intel quad-core processor.

- Contact: Bernard Maigret

6.7. Protein-Marshmallow

KEYWORDS: Coarse-Grained Representation

SCIENTIFIC DESCRIPTION

A Protein-protein interaction may be considered in terms of physical interaction between two deformable objects. The description at the atomic level of such complex objects is beginning to be feasible by MD simulations, but this requires the use of petaflop machines which are out of reach of most laboratories.

FUNCTIONAL DESCRIPTION

The Protein-Marshmallow program represents the surface of a protein as "coarse grained" 3D triangle mesh. In this mesh, each triangle is colored according to some biological property. In this way, a large complex object may be represented by a much smaller number of samples in a 3D mesh. The Marshmallow program describes deformations of such meshes under the influence of an external force field to simulate the strains that one object may undergo over time due to the interaction with another one.

- Contact: Bernard Maigret

6.8. Platforms

6.8.1. 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. 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 [19]. 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.

7.2. Large-Scale Analysis of 3D Protein Interactions

As part of a continuing collaboration with a former doctoral student in the Orpailleur team, Anisah Ghoorah (now at the University of Mauritius), we used her KBDOCK database of all known PPIs to perform a large-scale statistical analysis of the secondary structure composition of known protein-protein binding sites [14]. This showed that some combinations of secondary structure features are significantly favoured, whereas other combinations are considerably dis-favoured. These findings could provide knowledge-based rules for the prediction of unsolved protein-protein interactions.

7.3. Predicting Drug Side Effects

Together with Harmonic Pharma SAS (a LORIA / Inria spin-out company), we developed the "GESSE" method for proposing new uses for existing therapeutic drug molecules by associating the Gaussian shapes of known drug molecules with their clinically observed side-effects [15].

7.4. Modeling a GPCR Receptor Complex

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 [12]. 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.

7.5. Modeling the Apelin Receptor

The Apelin receptor (ApelinR) is a GPCR which is important in regulating cardiovascular homeostasis. 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 ApelinR [13]. This study provides new mechanistic insights which could lead to the development of therapeutic agents for the treatment of heart failure.

7.6. 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 the pathogenic fungus *Paracoccidioides lutzii* [17] which is a serious health threat, especially in Brasilian hospitals.

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

Participants: Marie-Dominique Devignes [contact person], Jérémie Bourseau.

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. IFB

Participant: Marie-Dominique Devignes [contact person].

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.

8.2.2.2. 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 choc 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 of International Scientists

8.4.1.1. Doctoral Students

In the frame of a collaboration with the University of Brasilia, Dr. A. Abadio and three doctoral students (A. Souza, J. Ribeiro, P. Alves) visited in July 2015.

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 is organising a workshop ("Atelier Santé") for the Fédération Charles Hermite.

9.1.1.2. Member of Organizing Committees

Marie-Dominique Devignes co-organised a workshop on Structural Modeling of Type IV Secretion Systems.

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-2015, NETTAB-2015, and MIVBM-2015.

9.1.2.2. Reviewer

David Ritchie was a reviewer for IJCAI-2015.

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, AIMS Biophysics, Bioinformatics, Current Opinion in Structural Biology, Journal of Biomedical Semantics, Journal of Chemical Information and Modeling, Journal of Molecular Modeling, Molecules, and Proteins.

9.1.4. Invited Talks

David Ritchie gave a presentation to the *Plateau de Modélisation Moléculaire Multi-échelle* (University of Reims).
Seyed Alborzi presented the EC-DomainMiner approach at *JOBIM-2015* (Clermont-Ferrand, France).
Bernard Maigret gave presentations for the *11th International Symposium on Bioinformatics Applied to Health* (State University of Maringá, Brasil) and the *Workshop Franco-Brasileiro de programa Ciência sem Fronteira* (EMPBRAPA, Brasil).
Marie-Dominique Devignes gave a presentation to the *Institute for Structural and Molecular Biology* at Birkbeck College (London, UK).

9.1.5. Scientific Expertise

David Ritchie has reviewed grant proposals for the French ANR and the British BBSRC.

9.1.6. Research Administration

Marie-Dominique Devignes is Chargée de Mission for the CyberBioSanté research axis at the LORIA.
David Ritchie is a member of the Commission de Mention Informatique (CMI) of the IAEM doctoral school of the University of Lorraine.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence: Marie-Dominique Devignes, *Relational Database Design and SQL*, 30 hours, L1, Univ. Lorraine.

Master: Marie-Dominique Devignes, *Biological Data Mining and Classification*, 12 hours, L3, Univ. Lorraine.

Doctorat: Bernard Maigret, *Virtual Screening*, 10-17 June, EMBRAPA, Brasil.

9.2.2. Supervision

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.

PhD in progress: Benoît Henry, *Probability theory applied to evolutionary biology,* 01/10/2013, Nicolas Champagnat, David Ritchie.

9.2.3. Juries

HdR: Julie Bernauer, *Méthodes géometriques et statistiques pour l'analyse et la prédiction des interactions structurales de biomolécules,* Université Paris Sud 11, 13/01/2015.

PhD: Romain Vasseur, *Développements HPC pour une nouvelle méthode de docking inverse : Application aux protéines matricielles,* Université de Reims – Champagne Ardennes, 29/01/2015, Pr Manuel Dauchez, Dr Stéphanie Baud, Dr Luiz-Angelo Steffenel.

PhD: Clovis Galiez, *Fragments structuraux : comparaison, prédictibilité à partir de la séquence et application à l'identification de protéines de virus,* Université de Rennes 1, 08/12/2015, Dr François Coste, Dr Jacques Nicolas.

PhD: Alicia Zhukov, *Knowledge-based generalization for metabolic models,* Université de Bordeaux, 18/12/2014, Dr David Sherman.

9.3. Popularization

An article on our KBDOCK software has been accepted for publication in ERCIM News (edition 104, January 2016) [18].

10. Bibliography

Major publications by the team in recent years

[1] A. 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

[2] A. 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

[3] A. 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

[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° 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] 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° 24, pp. 3274-3281 [*DOI :* 10.1093/BIOINFORMATICS/BTS618], https://hal.inria.fr/hal-00756813

[8] D. 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

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

[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] 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 p. [*DOI :* 10.1038/SREP10760], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01162594

[13] 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^{o} 1, pp. 314-322 [*DOI :* 10.1096/FJ.14-256339], https://hal.inria.fr/hal-01251633

[14] A. GHOORAH, M.-D. DEVIGNES, S. Z. ALBORZI, M. SMAÏL-TABBONE, D. RITCHIE. *A Structure-Based Classification and Analysis of Protein Domain Family Binding Sites and Their Interactions*, in "Biology", April 2015, vol. 4, n^{o} 2, pp. 327-343 [*DOI :* 10.3390/BIOLOGY4020327], https://hal.inria.fr/hal-01216748

[15] 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^{o} 9, pp. 1804-1823 [*DOI :* 10.1021/ACS.JCIM.5B00120], https://hal.inria.fr/hal-01216493

[16] 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 [*DOI :* 10.1107/S1600576715022931], https://hal.inria.fr/hal-01261402

[17] A. K. RODRIGUES ADABIO, E. S. KIOSHIMA, V. LEROUX, N. F. MARTINS, B. MAIGRET, M. S. SOARES FELIPE. *Identification of New Antifungal Compounds Targeting Thioredoxin Reductase of Paracoccidioides Genus*, in "PloS One", November 2015, vol. 10, n^{o} 11, e0142926 [*DOI :* 10.1371/JOURNAL.PONE.0142926], https://hal.inria.fr/hal-01251619

Articles in Non Peer-Reviewed Journals

[18] 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

[19] 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

References in notes

[20] 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

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

[22] 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

[23] 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

[24] 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

[25] M.-D. Devignes, B. Sidahmed, M. Smail-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

[26] S. E. Dobbins, V. I. Lesk, M. J. E. Sternberg. *Insights into protein flexibility: The relationship between normal modes and conformational change upon protein–protein docking*, in "Proceedings of National Academiy of Sciences", 2008, vol. 105, n^{o} 30, pp. 10390–10395

[27] D. Filmore. *It’s a GPCR world*, in "Modern Drug Discovery", 2004, vol. 7, pp. 24–28

[28] 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

[29] W. J. Frawley, G. Piatetsky-Shapiro, C. J. Matheus. *Knowledge Discovery in Databases: An Overview*, in "AI Magazine", 1992, vol. 13, pp. 57–70

[30] 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

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

[32] 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

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

[34] 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

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

[36] J. JANIN, K. HENRICK, J. MOULT, L. TEN EYCK, M. J. E. STERNBERG, S. VAJDA, I. VAKSER, S. J. WODAK. *CAPRI: A critical assessment of PRedicted Interactions*, in "Proteins", 2003, vol. 52, pp. 2–9

[37] 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

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

[39] 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.

[40] 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

[41] I. MOAL, P. 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

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

[43] 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

[44] 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

[45] 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

[46] 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, E. R. Helena Galhardas (editor), Lecture Notes in Computer Science, Springer, 2014, vol. 8574, pp. 16 - 31, https://hal.inria.fr/hal-01095591

[47] B. Pierce, W. Tong, Z. Weng. *M-ZDOCK: A Grid-Based Approach for C_n Symmetric Multimer Docking*, in "Bioinformatics", 2005, vol. 21, n$^{\mathrm{o}}$ 8, pp. 1472–1478

[48] 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

[49] 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$^{\mathrm{o}}$ 17, pp. 1865-1873 [*DOI :* 10.1093/bioinformatics/btn334], https://hal.inria.fr/inria-00434264

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

[51] 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

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

[53] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, H. J. Wolfson. *Geometry-based flexible and symmetric protein docking*, in "Proteins", 2005, vol. 60, n$^{\mathrm{o}}$ 2, pp. 224–231

[54] 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

[55] M. Smaïl-Tabbone. *Contribution to knowledge extraction from biological data*, Université de Lorraine, November 2014, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01093943

[56] 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

[57] A. B. Ward, A. Sali, I. A. Wilson. *Integrative Structural Biology*, in "Biochemistry", 2013, vol. 6122, pp. 913–915

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

[59] 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

[60] 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.