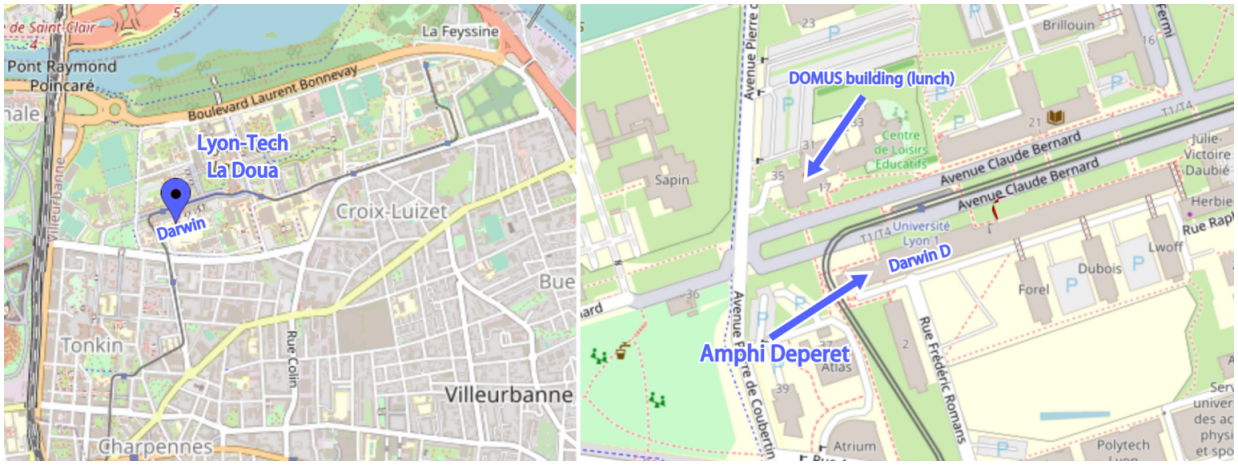
GT MASIM@Journées 2021 du GdR BIM

Université Claude Bernard, Lyon

Thursday 25th	
9:15-9:25	Welcome address
RNA session	
9:25-9:55	Vaitea Opuu – Max Planck Institute for Mathematics in the Sciences
	Efficient prediction of RNA folding pathways using the fast Fourier transform
10:00-10:30	Bruno Sargueil – CITCOM, Universite de Paris
	Investigating RNA tertiary structure using SHAPE chemical probing
10:30-11:00	Coffee break
11:00-11:30	Bertrand Marchand – LIX, Ecole Polytechnique; LIGM, Univ. Gustave Eiffel
	Tree Diet: Reducing the Treewidth to Unlock FPT Algorithms in RNA Bioinformatics
11:35-12:06	Guillaume Postic – IBISC, Univ. Paris-Saclay, Evry
	Deep learning techniques applied to RNA 3D structure prediction
12:00-14:00	Joint lunch with GT SeqBIM@DOMUS building (included in registration fees ;))
Docking and Complexes session	
14:00-14:30	Juliette Martin – Molecular Microbiology and Structural Biology UMR 5086 CNRS/Universite Lyon
	Contact stability and water in protein-protein interfaces studied by molecular dynamics simulations
14:35-15:05	Yasser M Behbahani – Sorbonne Universite
	DLA-Ranker: Evaluating protein docking conformations with many locally oriented cubes
15:10-15:40	Helene Bret – Institute for Integrative Biology of the Cell (I2BC), CEA
	Deep learning approach as a scoring method for rigid-body docking.
15:40-16:15	Coffee break
16:15-16:45	Stephane Teletchea – UFIP, Nantes Universite
	Structural Design and Analysis of the RHOA-ARHGEF1 Binding Mode: Challenges and Applications for Protein-Protein Interface Prediction
16:50-17:20	Mathilde Carpentier – Museum National d'Histoire Naturelle
	Protein folds as synapomorphies of the tree of life
Friday 26th	
9:00-9:15	Romain Launay – Toulouse Biotechnology Institute, TBI
	Characterization and functional comprehension of an enzymatic assembly: the Ubiquinone metabolon from Escherichia coli
9:15-9:45	Elin Teppa – Toulouse Biotechnology Institute, Universite de Toulouse, CNRS, INRAE, INSA
	Structural and sequence investigations of regioselectivity and substrate specificity in a family of enzymes: the case study of ubiquinone biosynthesis hydroxylases
Protein folding session	
9:50-10:20	Diego Zea – Independent researcher, Toulouse, France
	Impact of protein conformational diversity on AlphaFold2 predictions
10:20-11:00	Coffee break
11:00-11:30	Therese E Malliavin – Institut Pasteur, Universite de Paris
	Exploring exhaustively the conformational space of proteins using a discrete Distance Geometry approach
11:35-12:05	Frederic Cazals – ABS, Inria Sophia
	Mining protein flexibility: a new class of move sets based on Tripeptide Loop Closure

Location



The event will be hosted by Université Claude Bernard in Lyon, France.

Lectures will be delivered in the **Amphi Deperet**, located in the Darwin D building of the LyonTech – La Doua campus. Thursday's lunch is **included for registered participants**, and will be served in the DOMUS building, right across the street from the Darwin buildings.

Abstracts

Vaitea Opuu – Max Planck Institute for Mathematics in the Sciences

Efficient prediction of RNA folding pathways using the fast Fourier transform

Abstract: The function of non-coding RNAs is largely determined by their structure. For instance, ribozymes can often be analyzed in terms of basic structural motifs; other RNAs, like riboswitches, involve dynamical changes between alternative structures. Understanding the relation between sequence, structure, and dynamics is, therefore, a central challenge in molecular biology.

We investigated a simple two-steps approach to study RNA structure folding pathways: 1) The folding: We developed a folding algorithm inspired by the kinetic partitioning mechanism, by which molecules follow alternative folding pathways to their native structure, some much faster than others. Similarly, our algorithm RAFFT generates an ensemble of concurrent folding pathways ending in multiple metastable structures for each given sequence. RAFFT takes advantage of the fast Fourrier transform (FFT) to sample helices efficiently. 2) The folding kinetic: We proposed a kinetic folding ansatz, which is built with the concurrent folding pathways. From this kinetic ansatz, we obtained folding trajectories starting from the completely unfolded molecule to its set of metastable structures.

We first assessed the quality of the folding component with a well-curated benchmark dataset. Secondly, we used the proposed kinetic ansatz to reproduce folding trajectories. For the test cases we considered, our method was able to reproduce the folding kinetics qualitatively but using fewer structures.

This work has already shown some promising results in terms of speed, accuracy, and versatility. Our ongoing efforts aimed at 1) providing an efficient implementation, 2) diversifying the FFT-based helix sampling to other purposes. For example, one other use case is the sampling of RNA-RNA interaction sites.

References: Opuu, Merleau, Smerlak (2021) submitted; RAFFT: Efficient prediction of RNA folding pathways using the fast Fourier transform

Collaborators: Vaitea Opuu, Nono SC Merleau, Vincent Messow, and Matteo Smerlak

Bruno Sargueil – CITCOM, Universite de Paris

Investigating RNA tertiary structure using SHAPE chemical probing

Abstract: RNA structure modelling from chemical probing experiments has made tremendous progress, however accurately predicting large RNA structures is still challenging for several reasons. RNA are inherently flexible and often adopt many energetically similar structures, which are not reliably distinguished by the available, incomplete thermodynamic model. Moreover, computationally, the problem is aggravated by the relevance of pseudoknots and non-canonical base pairs, which are hardly predicted efficiently. To identify nucleotides involved in pseudoknots and non-canonical interactions, we scrutinized the SHAPE reactivity of each nucleotide of a benchmark RNA under multiple conditions, the results and their potential implication for RNA structure modeling will be presented.

Collaborators: Bruno Sargueil

Bertrand Marchand – LIX, Ecole Polytechnique; LIGM, Univ. Gustave Eiffel

Tree Diet: Reducing the Treewidth to Unlock FPT Algorithms in RNA Bioinformatics

Abstract: Hard graph problems are ubiquitous in Bioinformatics, inspiring the design of specialized Fixed-Parameter Tractable algorithms, many of which rely on a combination of tree-decomposition and dynamic programming. The time/space complexities of such approaches hinge critically on low values for the treewidth tw of the input graph. In order to extend their scope of applicability, we introduce the Tree-Diet problem, i.e. the removal of a minimal set of edges such that a given tree-decomposition can be slimmed down to a prescribed treewidth tw'. Our rationale is that the time gained thanks to a smaller treewidth in a parameterized algorithm compensates the extra post-processing needed to take deleted edges into account.
Our core result is an FPT dynamic programming algorithm for Tree-Diet, using $2^{O(tw)}n$ time and space. We complement this result with parameterized complexity lower-bounds for stronger variants (e.g., NP-hardness when tw' or $tw - tw'$ is constant). We propose a prototype implementation for our approach which we apply on difficult instances of selected RNA-based problems: RNA design, sequence-structure alignment, and search of pseudoknotted RNAs in genomes, revealing very encouraging results. This work paves the way for a wider adoption of tree-decomposition-based algorithms in Bioinformatics.

Collaborators: Bertrand Marchand, Yann Ponty and Laurent Bulteau

Guillaume Postic – IBISC, Univ. Paris-Saclay, Evry

Deep learning techniques applied to RNA 3D structure prediction

Abstract: The success of DeepMind AlphaFold (Jumper et al., 2021) at the CASP competition is arguably the major milestone on the way to solving the protein folding problem. Accurate predictions of protein structures are thus achieved by using deep learning techniques in combination with knowledge-based scoring functions. The latter are derived from statistics computed on experimentally determined protein structures. Here, we present the results obtained and the difficulty encountered when following an analogous approach, in our attempt to tackle the RNA folding problem. Thus, a convolutional and a recurrent neural network architectures (CNN and RNN) have been adapted to RNA structures to predict interatomic distances and pseudo-torsion angles. As the construction of a good training data set is critical before using deep learning algorithms, we had to build “RNANet” (Becquey et al., 2021), a database integrating information about RNA sequences, structures and interactions (https://evryrna.ibisc.univ-evry.fr/evryrna/rnanet). Finally, future developments at EvryRNA will also include RNA-specific statistical potentials.
References Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. Becquey, L., Angel, E., & Tahi, F. (2021). RNANet: an automatically built dual-source dataset integrating homologous sequences and RNA structures. Bioinformatics, 37(9), 1218-1224.

Collaborators: Fariza Tahi and Guillaume Postic

Juliette Martin – Molecular Microbiology and Structural Biology UMR 5086 CNRS/Universite Lyon

Contact stability and water in protein-protein interfaces studied by molecular dynamics simulations

Abstract: Proteins interaction are at the basis of many protein functions, and the knowledge of 3D structures protein-protein complexes provides structural, mechanical and dynamical support of these functions. Protein-protein interfaces can be seen as stable, organized regions where residues from different partners form non-covalent interactions that are responsible for interaction specificity and strength. They are commonly described as a peripheral region, whose role is to protect the core region, that concentrates the most contributing interactions, from the solvent. Here, we have carried out medium-range MD simulations (300 to 500ns) in explicit solvent on 8 different protein-protein complexes of different functional class and interface size. We show that in 6 out 8 cases, the interfaces rearrange during the simulation time, in stable and long-lived substates with alternative residue-residue contacts. These rearrangements are not restricted to side-chain fluctuations in the periphery but also affect the cores of interfaces, and up to 40% of the initial contacts. Another aspect of interfaces that is often neglected is the involvement of water molecules. Here we show that the number of water molecules that are displaced upon complex formation is an excellent indicator of interaction strength.

Collaborators: Elisa Frezza and Juliette Martin

Yasser M Behbahani – Sorbonne Universite

DLA-Ranker: Evaluating protein docking conformations with many locally oriented cubes

Abstract: Proteins ensure their biological functions by interacting with each other, and with other molecules. Determining the relative position and orientation of protein partners in a complex remains challenging. Here, we address the problem of ranking candidate complex conformations toward identifying near-native conformations. We propose a deep learning approach relying on a local representation of the protein interface with an explicit account of its geometry. We show that the method is able to recognize certain pattern distributions in specific locations of the interface. We compare and combine it with a physics-based scoring function and a statistical pair potential.

Collaborators: Yasser Mohseni Behbahani, Elodie Laine, Alessandra Carbone

Helene Bret – Institute for Integrative Biology of the Cell (I2BC), CEA

Deep learning approach as a scoring method for rigid-body docking.

Abstract: Protein complexes play a central role in the cell and determining the mechanisms of protein-protein interaction is essential for understanding most cellular processes. As the experimental determination of 3D complex structures is not always possible, protein docking methods aim to identify in silico the most likely conformations of the bound partners of a complex. Molecular docking is composed of two main steps, the generation of possible interface models (sampling step) followed by the scoring of these models in order to choose the most plausible ones. Several scoring methods have been proposed based on physics, statistical and evolutionary information as in the consensus score implemented in the server InterEvDock3 developed by our team. In parallel, deep learning approaches have proven to be extremely powerful to study the structure of biological objects, by extracting a signal from a covariation map as in ComplexContact (Zeng et al. NAR (2018)) or TrRosetta (Yang et al. PNAS (2020)), or more recently by analyzing protein sequences with the Transformers technology in the successful AlphaFold2 method (Jumper et al. Nature (2021)).
In this work we aim at developing a deep learning method for the rigid body docking scoring step. The goal is to directly score decoys by integrating physico-chemical context, geometry of the interfaces and evolutionary information if available. We will present the deep learning model that we have developed so far to train the network from monomer structures at the residue level. Inputs of our model include geometric and sequence information for one amino acid and its environment. The model uses a masking approach to evaluate the adequacy between the residue and its context. We aim to integrate this score over all amino acids involved in an interface model to be scored.

Collaborators: Helene Bret, Jessica Andreani, Raphael Guerois

Stephane Teletchea – UFIP, Nantes Universite

Structural Design and Analysis of the RHOA-ARHGEF1 Binding Mode: Challenges and Applications for Protein-Protein Interface Prediction

Abstract: The interaction between two proteins may involve local movements, such as small side-chains re-positioning or more global allosteric movements, such as domain rearrangement. We studied how one can build a precise and detailed protein-protein interface using existing protein-protein docking methods, and how it can be possible to enhance the initial structures using molecular dynamics simulations and data-driven human inspection. We present how this strategy was applied to the modeling of RHOA-ARHGEF1 interaction using similar complexes of RHOA bound to other members of the Rho guanine nucleotide exchange factor family for comparative assessment. In parallel, a more crude approach based on structural superimposition and molecular replacement was also assessed. Both models were then successfully refined using molecular dynamics simulations leading to protein structures where the major data from scientific literature could be recovered. We expect that the detailed strategy used in this work will prove useful for other protein-protein interface design. The RHOA-ARHGEF1 interface modeled here will be extremely useful for the design of inhibitors targeting this Protein-Protein Interaction (PPI).

Gheyouche E, Bagueneau M, Loirand G, Offmann B, Téletchéa S. Structural Design and Analysis of the RHOA-ARHGEF1 Binding Mode: Challenges and Applications for Protein-Protein Interface Prediction. Front Mol Biosci. 2021;8:643728. doi: 10.3389/fmolb.2021.643728

Collaborators: Stephane Teletchea

Mathilde Carpentier – Museum National d'Histoire Naturelle

Protein folds as synapomorphies of the tree of life

Abstract: To reconstruct the Tree of Life, protein and DNA sequences have been used for long, but a level of protein 3D structural organization has been neglected, fold. Folds are defined as the architecture and topology of secondary structures. 1,232 of these have been characterized. This article aims to explore whether such folds could provide reliable synapomorphies for some clades of the tree of life. We map folds onto a tree of life and measure the consistency of each fold character. As a result, 20% of the folds are present in all superkingdoms, and 53.9% are potential synapomorphies. We find fold characters consistently supporting several nested eukaryotic clades with divergence times spanning from 1,100 mya to 380 mya. As for the earliest branches of the tree of life, the three superkingdoms are discriminated by eukaryotic specific folds (181) as well as shared folds between Eukaryota and one of the two other superkingdoms. Many folds shared by parts of eukaryotes and some eubacteria should result from past horizontal transfers (e.g. cyanobacteria to photosynthetic eukaryotes) witnessing significant fold flow to eukaryotes. Among eukaryotes, some folds therefore appear as synapomorphies of the species phylogeny, while others are markers of transfers to Eukaryota.

Collaborators: Mathilde Carpentier

Romain Launay – Toulouse Biotechnology Institute, TBI

Characterization and functional comprehension of an enzymatic assembly: the Ubiquinone metabolon from Escherichia coli

Abstract: Protein-protein interactions and supramolecular complexes are essential for the functioning of living cells. They play a huge role in a number of biological functions, such as signal transduction, cell-to-cell communication, transcription, replication, and membrane transport. Determination and characterization of such interactions remain a challenge for structural biology. However, the progress in the development of computational methods and the powerful computing resources available nowadays have enabled considerable improvement in the accuracy of 3D-molecular assembly predictions. In this context, our work focuses on a protein complex from Escherichia coli involved in the ubiquinone biosynthesis. Ubiquinone is a redox-active prenol lipid, composed of two parts, i.e. a hydrophobic tail

(isoprene unit) and a redox active head (Figure 1A). Due to its crucial role in the cell, the latter is present in most of the organisms and mainly located in the cell membranes (Stefely and Pagliarini, 2017). This molecule has many key conserved molecular functions, notably as electron shuttle in the mitochondrial respiratory chain and as antioxidant for the reduction of hydroxyl radicals. Ubiquinone and its functions are very conserved across the species, making this molecule very interesting to investigate.

Collaborators: Romain Launay, Elin Teppa, Carla Martins, Sophie Abby, Fabien Pierrel, Isabelle Andre, Jeremy Esque

Elin Teppa – Toulouse Biotechnology Institute, Universite de Toulouse, CNRS, INRAE, INSA

Structural and sequence investigations of regioselectivity and substrate specificity in a family of enzymes: the case study of ubiquinone biosynthesis hydroxylases

Abstract: The ability of enzymes to selectively catalyze the modifications of compounds is essential for the proper functioning of most biological pathways. Some enzymes present a broad substrate specificity, being able to act upon a group of similar substrates, whereas others present a remarkable ability to catalyze the reaction with a specific compound. Identification of amino acid residues responsible for the specificity helps to understand the enzyme's molecular mechanism and opens the possibility of fine-tuning this natural function for eventually biotechnological or therapeutic applications. Besides substrate specificity, some enzymes may catalyze a reaction in a regioselective manner, which refers to the preference of a chemical bond formation/breaking at one particular position over all the other possible positions. Here we present a computational approach applied to characterize the active site of a family of enzymes showing variation in their specificity and regioselectivity. We aim to understand the structural and molecular bases governing the differences in substrate specificity and the determinants of the narrow/broad regioselectivity in a family of enzymes involved in the biosynthesis of ubiquinone.

Collaborators: Elin Teppa, Romain Launay,Alexandre G de Brevern, William Schmitt, Ivan Junier, Sophie Abby, Fabien Pierrel, Jeremy Esque, Isabelle Andre

Diego Zea – Independent researcher, Toulouse, France

Impact of protein conformational diversity on AlphaFold2 predictions

Abstract: The outstanding breakthrough of AlphaFold2 on the prediction of tridimensional protein structures has impacted the field of structural bioinformatics. While the software was tested on the resolution of single conformations, we wonder whether the tool can predict different structural conformationsof a protein. The prediction of structural diversity could improve our knowledge of the dynamics and functional mechanisms of the modeled proteins.
Due to the dynamical nature of protein structure, the native state of a protein is composed of an ensemble of structural conformations. In this work, we analyzed protein structures that have been experimentally resolved in the presence and absence of the biological ligand as a sample from that ensemble. We curated by hand the data set to ensure that we have meaningful conformational changes between the different conformations. We end up with a set of pairs of apo (unbound) and holo (bound) forms for 95 proteins, which we used to challenge the AlphaFold2 method to test its ability to predict both conformations.
By measuring the RMSD between the models and the known apo and hold forms, we found that for 70% of our proteins, AlphaFold2 predicts only the holo form. Therefore, it is unable to predict both conformations even when multiple models are requested. However, AlphaFold2's tendency to model the conformation bound to the biological ligand makes it an exciting tool for analyzing protein-ligand interactions or performing docking experiments.
More importantly, we found that AlphaFold2 decreasesin performance as the conformational diversity of the protein increases. This impairment is related to the heterogeneity in the degree of conformational diversity found between different protein family members. Here, we show that AlphaFold2 has a lower performance on proteins showing high conformational diversity, i.e., significant structural differences between their conformations. What's more, we observed the same tendency for proteins belonging to a family whose members show a large conformational diversity. As AlphaFold2 takes a multiple sequence alignment as input, we hypothesize that the evolutionary information cofounded by the different structural ensembles of the family members leads to that diminished performance.

Collaborators: Tadeo Saldano, Nahuel Escobedo, Julia Marchetti, Diego Javier Zea, Juan Mac Donagh, Ana Julia Velez Rueda, Eduardo Gonik, Agustina Garcia Melani, Julieta Novomisky Nechcoff, Marin N. Salas, Tomás Peters, Nicolas Demitroff, Sebastian Fernandez Alberti, Nicolas Palopoli, Maria Silvina Fornasari, Gustavo Parisi

Therese E Malliavin – Institut Pasteur, Universite de Paris

Exploring exhaustively the conformational space of proteins using a discrete Distance Geometry approach

Abstract: The optimization problem encountered in protein structure determination is undergoing a change of perspective due to the larger importance in biology taken by disordered regions of biomolecules. In such cases, the convergence criterion is more difficult to set up; moreover, the enormous size of the space makes it difficult to achieve a complete exploration. The interval Branch-and-Prune (iBP) approach, based on a reformulating of the Distance Geometry Problem (DGP) provides a theoretical frame for the exhaustive sampling of the conformations.
An implementation of the iBP approach, oriented toward the sampling of protein structure, was recently proposed (Worley et al, 2018; Malliavin et al, 2019). Basing on this approach, a pipeline was applied on a partially disordered linker (Malliavin, 2021) using Nuclear Magnetic Resonance (NMR) chemical shifts and small angle X-ray scattering (SAXS) to determine representative conformations and their corresponding populations. The perspectives of this work for the study of biomolecular conformations will be discussed.
References
Worley B, Delhommel F, Cordier F, Malliavin T, Bardiaux B, Wolff N, Nilges M, Lavor C, Liberti L. Tuning interval Branch-and-Prune for protein structure determination. J Glob Optim 72, 109 (2018).
Malliavin TE, Mucherino A, Lavor C, Liberti L. Systematic Exploration of Protein Conformational Space Using a Distance Geometry Approach. J Chem Inf Model 59, 4486 (2019).
Malliavin TE. Tandem domain structure determination based on a systematic enumeration of conformations. Scientific Reports volume 11, Article number: 16925 (2021)

Collaborators: Therese E Malliavin

Frederic Cazals – ABS, Inria Sophia

Mining protein flexibility: a new class of move sets based on Tripeptide Loop Closure

Abstract: Protein flexibility is key for most biological functions, yet, poses daunting challenges from the computational standpoint. On the one hand, the prediction of large amplitude conformational changes requires exploring conformational spaces with (tens of) thousands of degrees of freedom. On the other hand, localized dynamics fine tuning thermodynamics and kinetics are equally challenging to predict accurately. To complement structural approaches targeting the prediction of static structures, methods unveiling these complexes dynamics are called for.

At the heart of protein flexibility studies, are so-called move sets, that is operations generating novel protein conformations from an existing one. This talk will discuss two aspects in this realm, related to the celebrated Tripeptide Loop Closure (1,2,3). First, starting with the review of the Ramachandran diagram, we will discuss properties of backbone reconstructions yielded by TLC, showing that TLC provide a much more thorough sampling of Ramachandran space (4). Second, we will present a novel class of move sets, based on a global parameterization of the backbone of a polypeptide chain (5).
References
(1) Evangelos A Coutsias, Chaok Seok, Matthew P Jacobson, and Ken A Dill. A kinematic view of loop closure. Journal of computational chemistry, 25(4):510–528, 2004.
(2) D.J. Mandell, E.A. Coutsias, and T. Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nature Methods, 6:551-552, 2009.
(3) A. Barozet, K. Molloy, M. Vaisset, T. Simeon, and J. Cortes. A reinforcement-learning-based approach to enhance exhaustive protein loop sampling. Bioinformatics, 2019.
(4) T. O'Donnell, C. Robert and F. Cazals, Tripeptide loop closure: a detailed study of reconstructions based on Ramachandran distributions, under revision.

(5) T. O'Donnell, and F. Cazals, Protein loops sampling based on a global parameterization of the conformational space, In preparation.

Collaborators: T. O'Donnell, Frederic Cazals