# PROTEIN AND NUCLEOTIDE STRUCTURE

Chairs: Anna Tramontano and Jan Gorodkin

# C-1. Developing a methodology for protein-ligand docking based on genetic algorithm and normal modes

*Lima AN (2), Philot EA (2), Scott LPB (1,*), Perahia D (2)*

The inconvenience of the fully flexible docking is the high computational cost. Consequently, few method. Different approximations exist depending which type of degrees of freedom are to be considered , and this includes rigid docking where protein and ligand are rigid, semi-flexible where only a selected set of side-chains of the protein are allowed to move, or fully flexible where all the protein (including the backbone) and the ligand are allowed to change their conformation. A methodology that allows a fully flexible docking is very interisting.

## *Materials and Methods*

GANM combines Genetic Algorithm and Normal Modes and uses a amino-acid Rotamer Library to simulate protein-ligand docking. Within the framework of GA the degrees of freedom used include the rigid translational and rotational motions of the ligand as well as its internal flexibility, the rotameric states of protein side-chains within and around the binding cavity, and the lowest frequency modes of the protein that describe its global conformational changes. We are implementing and testing the methodology in four versions.

## *Results*

We present here the docking results obtained with the application of this two first version of GANM to different protein-ligand syste and we compare our results wiht the reuslts of ohters docking programs.

## *Discussion*

We present the GANM method using a rotamer library to simulate “semi-flexible” protein-ligand docking which consider a rigid backbone and a rigid ligand but flexible side chains and the second version with the backbone semi-flexible. Although this two first versions do not treat the protein as being completely flexible, it can however be interesting to use it as a first approach to rapidly dock a ligand to a protein or to achieve a virtual screening for searching putative ligands.

## *Presenting Author*

Luis P. B. Scott (luis.scott@ufabc.edu.br)
UFABC - CMCC

## *Author Affiliations*

(1) Universidade Federal do ABC, CMCC. (2)Universidade Federal do ABC , Program of Master in Information Eginnering (3)Université Paris-Sud, Orsay, France

## *Acknowledgements*

The authors aknowledge the Fapesp (Brazil) and the CNRS (France) for the support to this work

## C-2. A statistical potential for modelling of protein-RNA complexes

*Tuszynska I (1,2*), Rother K (1,3), Bujnicki JM (1,3)*

RNA-protein interactions are extremely essential to gene expression in the cell. Methods which are able to predict protein-RNA complexes are needed to understand the principles of protein-RNA recognition and next to design new RNA-binding proteins.There are two statistical potentials for protein-RNA interaction prediction, which are developed by the Varani and the Fernandez groups. But first potential recognised very near native protein-RNA structures, with RMS

### *Materials and Methods*

Our potentials comprise a distance-dependent energy term (Er),angular-dependent energy component (Ea),site-dependent energy term (Es) and penalty for sterical clashes (Ep). Before the calculation, potentials reduce the all atom representation to the reduced representation, which is used in the Refiner program for a protein and the RedRNA program for RNA developed in our laboratory.

### *Results*

Score-rmsd plots as results of scoring of four decoys from bound docking. There are two testing sets generated by our group with GRAMM program and by the Varani group. Each test set was scored by our quasichemical potential and the potential developed by the Varani group. The first best scored decoy of our quasichemical potential function for each complex from the testing set compared to the native complex are presented. Only proteins are superimposed.

### *Discussion*

1.Our potentials discriminate native-like (with RMS < 10 Å) structures of protein-RNA complexes, while the potential developed by the Varani group recognises structures very close to the native (RMS < 5 Å) 2.Potentials give better results for thetest setobtained bytheVarani group, where perturbation was applied to the native structures of complexes. Hence, the using of refinement for GRAMM rigid body docking decoys before the application ofour potentials may improve the discrimination power of potentials. 3.To eliminate false positives in the final stage of the prediction of native structure o

### *Presenting Author*

Irina Tuszynska (irena@genesilico.pl)
International Institute of Molecular and Cell Biology

### *Author Affiliations*

1. Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland 2. Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland 3. Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, 61-614 Poznan, Poland

### *Acknowledgements*

This work was supported by following grants:HISZPANIA/152/2006from the Polish Ministry of Scienceand the EU 6FP Network of Excellence EURASNET (grant LSHG-CT-2005-518238)

## C-3. ProBiS: a web server for detection of structurally similar protein binding sites

*Konc J (1,*), Janežič D (1,2)*

Exploitation of locally conserved three-dimensional patterns of physicochemical properties on the surface of a protein for detection of binding sites that may lack sequence and global structural conservation.

### *Materials and Methods*

Web server ProBiS compares the query protein to members of a database of protein 3D structures and detects with sub-residue precision, structurally similar sites as patterns of physicochemical properties on the protein surface. Using an efficient maximum clique algorithm, the server identifies proteins which share local structural similarities with the query protein and generates structure-based alignments of these proteins with the query. Structural similarity scores are calculated for the query protein's surface residues, and are expressed as different colors on the query protein surface.

### *Results*

Given a structure of a protein with unknown binding sites, ProBiS suggests the regions on its surface which may be involved in binding with small ligands, proteins or DNA/RNA. Alternatively, given a protein with an identified binding site, ProBiS finds other proteins with structurally or physicochemically similar binding sites. If used as a pairwise structure alignment program, ProBiS detects and superimposes similar functional sites in a pair of submitted protein structures, even when these do not have similar folds.

### *Discussion*

Detection of binding sites which depends on structural similarity of protein surfaces is useful and accurate and can enjoy success in structures with dissimilar folding patterns. Structural similarity can provide additional advantages over sequence conservation in the detection of functional regions such as binding sites. Detailed instructions and user guidelines for use of ProBiS web server are available at http://probis.cmm.ki.si under 'HELP' and selected examples are provided under 'EXAMPLES'.

### *URL*

*http://probis.cmm.ki.si*

### *Presenting Author*

Janez Konc (konc@cmm.ki.si)
National Institute of Chemistry

### *Author Affiliations*

(1) National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia (2) University of Primorska, Faculty for Mathematics, Natural Sciences and Information Technologies, Glagoljaška 8, 6000 Koper, Slovenia

### *Acknowledgements*

Ministry of Higher Education, Science and Technology of Slovenia; Slovenian research agency (Z1-31666).

## C-4. Toward the prediction of the absolute quality of single protein structure models

*Benkert P (1,2,*), Biasini M (1,2), Schwede T (1,2)*

Estimating the quality of predicted structural models is a vital step in homology modeling since the individual models may contain considerable errors. To identify such inaccuracies, scoring functions have been developed which analyse different structural features of the protein models in order to generate a quality estimate. Most state-of-the-art scoring functions are primarily designed to rank alternative models of the same sequence. However, the relative quality of a model within an ensemble of models is not sufficient for determining its usefulness for different biomedical applications.

### *Materials and Methods*

Recently, we have introduced the composite scoring function QMEAN which consists of four statistical potential terms (two non-bonded interaction terms, a torsion angle term and a solvation term) and two components describing the agreement between predicted and observed secondary structure and solvent accessibility. QMEAN has been successfully used in model selection and has now been extended to provide an estimate of the absolute quality of a single model. This QMEAN Z-score is a measure of the likelihood that a given model is of a quality comparable to experimental structures.

### *Results*

We calculated the QMEAN Z-score for all protein chains from the PDB solved by X-ray crystallography and identified outliers with very high/low score. On the one side of the spectrum we observe positive QMEAN Z-score > 3 standard deviations for extremely stable proteins from (hyper-)thermophilic organisms. On the other side, membrane proteins which are known to require a lipid bilayer to maintain their structural integrity, result in negative QMEAN Z-score > 3 standard deviations. We also identified several structures in the PDB database with very low Z-scores, some known to be fabricated.

### *Discussion*

These results indicate that the new QMEAN score is reasonable measure of the stability and integrity of protein structures. We show that the QMEAN Z-score can not only be used to identify experimental structures of poor quality but also to assess the quality of theoretical protein models and thereby serve as an absolute measure of model quality.

### *URL*

http://swissmodel.expasy.org/qmean

### *Presenting Author*

Pascal Benkert (pascal.benkert@unibas.ch)
Swiss Institute of Bioinformatics, Biozentrum, University of Basel

### *Author Affiliations*

1 Biozentrum, University of Basel, Switzerland 2 SIB Swiss Institute of Bioinformatics, Basel, Switzerland

### *Acknowledgements*

SIB Swiss Institute of Bioinformatics

## C-5. NMR data and protein structure

*Vranken W (1,*)*

Nuclear Magnetic Resonance (NMR) is one of the key experimental techniques to determine the structure of biomolecules. Compared to X-ray crystallography, NMR data is complex and the software to analyse it fragmented. At the PDBe we have been working with the Collaborative Computing Project for the NMR community (CCPN) framework to make archived NMR data consistent with archived Protein Data Bank (PDB) coordinate data. We make this extensive data available to the community and have started to analyse it to gain new insights into the relation between protein structure and NMR paramaters.

### *Materials and Methods*

All software is developed in Python. The CCPN framework is used to store all NMR data, the CcpNmr FormatConverter to parse all archived data files. The NMR data is analysed and graphs are generated with the RPy package (connecting Python to R). Django is used for dynamic display of the web pages.

### *Results*

1. We now capture as much NMR data as possible at the time the coordinates are deposited by accepting full CCPN projects. 2. We continue to provide reference data for CCPN and, in collaboration, consistent NMR restraints and recalculated NMR structure coordinates. 3. We develop software for the community using the CCPN framework, for example for NMR data format data conversions (CcpNmr FormatConverter). 4. We collaborate to relate chemical shift and NMR restraint data to coordinate information, and work on analysing NMR structure ensembles.

### *Discussion*

Researchers who want to use structures determined by NMR often find it difficult to make sense of the way the structures are presented, or have difficulty locating the information they require in a standardized form. At the PDBe, we continue to work with the NMR community to provide services and analyses to help the user and make NMR data more accessible.

### *URL*

http://www.pdbe.org/nmr

### *Presenting Author*

Wim F. Vranken (wim@ebi.ac.uk)
European Bioinformatics Institute

### *Author Affiliations*

Protein Data Bank in Europe (PDBe) European Bioinformatics Institute (EBI)

## C-6. Predicting the effect of mutations on the thermal stability of proteins with statistical potentials and artificial neural networks

*Folch B (1), Rooman M (1), Dehouck Y (1,*)*

Proteins are exploited in applications such as the design of therapeutic agents, or in the agro-food or biotechnology sector. To improve yield or specificity, it is interesting to tune their properties via amino acid substitutions. In particular, maintaining its activity in unusual temperature conditions is often crucial. Although several methods have been designed to predict changes in thermodynamic stability at room temperature upon mutation (DDG), there is still a stringent need for predictive models that focus specifically on the effect of mutations on the thermal resistance of proteins.

### *Materials and Methods*

Our method relies on a set of statistical potentials, extracted protein structures, describing the couplings between four protein descriptors (sequence, distance, torsion angles and solvent accessibility). Terms accounting for volume variations upon mutation are also considered. The change in melting temperature is expressed as a linear combination of these potentials, whose weights are functions of the solvent accessibility of the mutated residue, identified with the help of a neural network. The network is trained and validated on a dataset on 1601 experimentally characterized mutations.

### *Results*

We used a 5-fold validation procedure to assess the performances of our model. The correlation coefficient (R) between computed and measured changes in melting temperature (DTm) is 0.73, for 90% of the mutations. In contrast, when DDG predictions are used to estimate DTm values, the performances drop significantly (R=0.67). Interesting differences are observed between DTm and DDG predictors. In particular, to predict thermal stability changes, local interactions are more important in the core than on the surface, although the opposite trend is observed for DDG predictions.

### *Discussion*

We showed that accurate predictions of thermal stability changes (DTm) upon mutation can only be achieved through predictive models specifically focused on thermal stability. Indeed, even though thermodynamic stability changes (DDG) are related to thermal stability changes, DDG predictions only allow a poor evaluation of DTm values. Moreover, our model reveals that for some types of interactions, the weighting functions strongly differ between DDG and DTm predictors. This indicates differences between the mechanisms that rule protein stability at room temperature and at higher temperatures.

### *Presenting Author*

Yves Dehouck (ydehouck@ulb.ac.be)
Université Libre de Bruxelles

### *Author Affiliations*

(1) Unité de Bioinformatique génomique et structurale, Université Libre de Bruxelles, Belgium.

### *Acknowledgements*

We acknowledge support from the Belgian State Science Policy Office through an Interuniversity Attraction Poles Programme (DYSCO), the Belgian Fund for Scientific Research (FNRS) (FRIA grant to BF and FRFC project), the Héger-Massa Fund and the Brussels Region (TheraVip project). MR is Research Director at the FNRS.

## C-7. BioShell: a universal utility library for structural bioinformatics

*Gront D (*)*

The amount of biological data can be overwhelming, therefore automated processing methods are indispensable in bioinformatical research. BioShell package is one of the very few tools that cover very wide range of applications, providing keeping intuitive design and great ease of use.

### *Materials and Methods*

BioShell project has been started in 2005 as a set of stand-alone programs. It has since then evolved to a fully featured scripting language for biomolecular modeling and structural bioinformatics. It is addressed to a wide audience of users. It can be used in three ways: • by calling command-line programs to do simple tasks as statistical data analysis, file parsing, sequence and structure alignment and rms calculations • as a library of modules for scripting languages running on JVM, such as Python or Ruby implementations • as a library for bioinformatics software development in Java.

### *Results*

BioShell has already played the crucial role in several scientific projects. For example, its graph library combined with a PDB parser were used to scan the whole PDB to find a particular topology of interdigitated b-strands. In a another case a large set of binding pockets was automatically detected and compared. Other applications comprises various calculations on biomolecular structures and processing trajectories from molecular simulations. The newest features include clustering of protein structures, optimal structure alignment (for any arbitrary coverage) and a PsiBlast parser/filter.

### *Discussion*

One of the biggest challenges in designing a computational tool is to combine the variety of different modules and options with the ease to learn and to use. BioShell users may begin with simple command-line tools and later switch to writing scripts. The combination of Java with Jython gives an easy access to any fragment of the BioShell code. The project's website provides many examples that cover almost the whole range of BioShell applications.

### *URL*

*http://bioshell.chem.uw.edu.pl/*

### *Presenting Author*

Dominik Gront (dgront@gmail.com)
University of Warsaw, Faculty of Chemistry

### *Author Affiliations*

University of Warsaw, Faculty of Chemistry ul. Pasteura 1, 02-093 Warsaw, Poland

### *Acknowledgements*

Support from Marie Curie fellowship (FP7-people-IOF) is greatly acknowledged.

## C-8. Molecular dynamic simulation studies of membrane bound fully solvated β3 Adrenergic Receptor

*Tewatia P (1, *), Malik BK (1), Sahi S (2)*

β3-adrenergic receptors (β3- ARs) belong to the widely studied class A of the G-protein coupled receptor (GPCR) super family. β3-ARs are located in the plasma membrane of both white and brown adipocytes where they mediate metabolic effects such as lipolysis and thermogenesis. They are reported to exist in human heart, gastrointestinal tract, urinary bladder detrusor, prostate, brain and in near term myometrium. Researchers are working to understand their function and regulation in different human tissues as they are also attractive drug target for treatment of obesity & adult-onset diabetes.

### *Materials and Methods*

A 5.0 ns Molecular dynamics (MD) simulation was performed of the 3D-model of β3-AR using the DESMOND version 2.2. The 3D-model of β3 AR was solvated with three point simple point charge (SPC) water molecules, 1-palmitoyl-2-oleoyl-phosphatidylcholine (POPC) membrane was also incorporated. The system was subjected to 20,000 steps of conjugate gradient energy minimization. Simulation was performed in the NVT ensemble with thermostat set at 300 K. Conformations were saved every 10 ps. All calculations were performed on a Linux cluster of 8 computers with 16 processors.

### *Results*

Comparing the initial structure and the structure after 5.0 ns simulation dynamic perturbations in the structure of the protein especially in the loop regions were observed. Matching the Cα of the transmembrane helices give RMSD of 3.2 Å and for the loops the Cα RMSD is 6.8 Å. The loop regions: I1 (intracellular loop 1 residues 64 to 72), I2 (residues 134 to 155), I3 (residues 226 to 292), E1 (Extracellular loop 1) and some regions of the helices showed major dynamic perturbations. These were further analyzed as I2 and I3 loops are important for selectivity and affinity for G proteins.

### *Discussion*

Comparing the initial structure to the structure after 5 ns simulation gives RMSD of 3.85 Å with major conformational changes in the loop regions. The helices are tilted a little closer to each other. On superimposing and comparing the trajectories obtained during production phase an average RMSD of 1.3Å was observed indicating fair overall stability of the different conformations in the helical region and high flexibility in the loop regions. The trajectory was analyzed and the average coordinate file generated during production phase of simulation was taken as the final model of β3 AR.

### *Presenting Author*

Parul Tewatia (ptewatia@amity.edu)
Amity Institute of Biotechnology, Amity University, Noida, UP, India

### *Author Affiliations*

(1)Amity Institute of Biotechnology, Amity University, Expressway, Sec 125, Noida, UP, India (2) School of Biotechnology, Gautam Buddha University, Greater Noida, UP, India

## C-9. The significance of the ProtDeform score

*Rocha J(*), Alberich R*

When a researcher uses a program to align two proteins and gets a score, one of her main concerns is how often the program gives a similar score to pairs that are or are not in the same fold. This issue was analysed in detail recently for the program TM-align with its associated TM-score. It was shown that because the TM-score is length independent, it allows a P-value and a hit probability to be defined depending only on the score. Also, they found that the TM-scores of gapless alignments closely follow an Extreme Value Distribution. ProtDeform score (PDscore) needs to be analysed.

### *Materials and Methods*

In this study, we define the PD-score in a coarser version based on backbone fragments instead of amino acids (Bioinformatics, 25: 1625-1631, 2009). We use the Gold Standard Benchmark (BMC Structural Biology, 9:23, 2009) and the one used for the TM-score analysis and one for gapless score analysis. We fit PD-scores frequencies to EVD and show PD-score independence from protein length. We calculate the a posteriori probability of a pair in the same fold given a PD-score and found the score thresholds also for homology discrimination. We plot prob. of Type I errors vs prob. of Type II errors.

### *Results*

This study on the ProtDeform score reveals that it is also length independent and that PD-scores of gapless alignments also follow approximately an EVD. Using the Gold Standard benchmark, PD-scores have lower probabilities of error than TM-scores all at a similar speed. The analysis is extended to homology discrimination showing that, again, ProtDeform gives higher hit probabilities than TM-align. We suggest using three different P-values according to the three different contexts: Gapless alignments, optimised alignments for fold discrimination and that for superfamily discrimination.

### *Discussion*

We estimated three different P-values for the three different discrimination problems we are faced with: one is for the scores obtained for gapless alignments between domains of the same length; a second one is for the scores produced by optimisation programs on domains of different topology; and the third one is for the scores produced by the same programs on domains of different homology. Mixing these hypotheses can under-or over-estimate the significance of the scores seen. We have found that PD-scores are length independent, discriminant and with a known significance.

### *URL*

*http://bioinfo.uib.es/~recerca/ProtDeform/*

### *Presenting Author*

Jairo E. Rocha (jairo@uib.es)
University of the Balearic Islands

### *Author Affiliations*

Department of Mathematics and Computer Science and University Institute for Health Research University of the Balearic Islands

### *Acknowledgements*

Spanish Ministry of Science and Technology [MTM2009-07165].

## C-10. Exploiting synergy between computational biology and X-ray crystallography for solving challenging macromolecular structures

*Wiegels T (1,*), Lamzin V (1)*

Solving structures of large macromolecular complexes is a challenging task in macromolecular crystallography (MX). Such crystals rarely diffract to high resolution, resulting in a scarce amount of data for generating electron density maps. Automated modelling approaches have been focused on high-resolution and their application to data extended to less than 3.0 Å typically results in incomplete and highly fragmented models. Therefore a use of complementary information for structure completion is highly desirable for robust and automated solution of low-resolution MX structures.

### *Materials and Methods*

ARP/wARP - one of the leading MX software projects - uses small protein-specific building blocks (residues or short poly-peptides) to build a model from MX electron density maps in an iterative approach. At least 50% of all protein structures in the PDB contain NCS-related fragments (subunits or whole molecules). By using an all-vs-all least squares superposition of the built fragments, we identify NCS relations. Typically, fragments are built differently in different NCS-related copies. Thus the identified NCS relations are used to extend and connect fragments in subsequent building cycles.

### *Results*

The method has been successfully tested on several cases. In a proof of principle 30% of residues of a homodimer (pdbID 1c48:A,B) were artificially removed. Starting from this 70% complete model at 1.6Å resolution ARP/wARP improved it to 78% (still highly fragmented) where only 50% of the sequence could be docked. However, by the aid of the NCS-based fragment extension, the completeness was improved to 96% and 93% of the sequence could be docked already after two building cycles. We will also present results for a large-scale test with data from our remote web-based MX computational facility.

### *Discussion*

The use of NCS provides a significant improvement in many cases, often in less model-building cycles. Since 50% of all structures contain NCS, the presented development may become an important contribution in the automated derivation of 3D structures from low-resolution MX data. The possibility to determine more structures in an automated manner would be a valuable asset to biomedical and pharmaceutical research.

### *URL*

*http://www.embl-hamburg.de/research/unit/lamzin/index.html*

### *Presenting Author*

Tim Wiegels (wiegels@embl-hamburg.de)
EMBL, Hamburg Outstation

### *Author Affiliations*

(1) European Molecular Biology Laboratory (EMBL), Hamburg Outstation

### *Acknowledgements*

EMBL International PhD Programme (EIPP)

## C-11. Pattern recognition with moment invariants for interpretation of very low resolution macromolecular density maps

*Heuser P*, Lamzin VS*

3D structural studies of macromolecular complexes often yield data to only very low resolution (cryo EM/X-ray Crystallography). The interpretation of such data usually starts with the segmentation of the map (e.g., with Watershed algorithm), which does not always give satisfactory results (over-segmentation/incorrect structural borders). Overall, fitting of known structures currently requires a lot of human expert knowledge and interaction so that an automated procedure is a highly desirable.

### *Materials and Methods*

We use 3rd order moment invariants to identify regions in density maps of macromolecular complexes that match the corresponding regions in the known structures of the constituting fragments. False positives are eliminated using difference distance matrices. Finally the structures of the fragments are placed into the map. Third order moment invariants give a concise but comprehensive description of 3D objects in only 11 numerical values, providing convenient means for fast search through a large amount of 3D data.

### *Results*

The method has been tested on calculated structure factors for large macromolecular complexes (genotoxin and GroEL) with 10 or 15 Å high-resolution limit. The individual subunits were fitted in the low-resolution density maps with an average r.m.s.d. on Cα atoms of 1.3 Å. New results obtained for experimental EM data will also be presented.

### *Discussion*

Since the last decade there have been many attempts to develop reliable 3D map segmentation algorithms with varying success, in order to reduce the complexity of the challenging task of low-resolution density map interpretation. The method presented here does not require a map segmentation step and provides accurate results without human interaction in reasonable time, due to the use of sophisticated pattern recognition algorithms. Implementation of real-space refinement procedures is expected to improve the results even further.

### *URL*

*http://www.embl-hamburg.de/research/unit/lamzin/index.html*

### *Presenting Author*

Philipp Heuser (philipp.heuser@embl-hamburg.de)
EMBL-Hamburg

### *Author Affiliations*

EMBL-Hamburg

### *Acknowledgements*

EMBL for a postdoctoral fellowship

## C-12. PTools: an open source molecular docking library

*Poulain P (1,*), Saladin A (2), Fiorucci S (3), Zacharias M (4), Prévost C (5)*

Most biological processes in the cell involve macromolecules interacting with one or several partners. Knowledge of the overall structures of these assemblies as well as the details of the interactions is essential for understanding the underlying biological mechanisms. Macromolecular docking is a challenging field of bioinformatics that tackles this question. PTools is an open source library dedicated to molecular docking simulations.

### *Materials and Methods*

PTools is an object-oriented Python/C++ library that contains low-level routines like PDB-format manipulation functions as well as high-level tools to perform docking simulations and analyze results. PTools deals with reduced (coarse-grained) representation for proteins and DNAs allowing simulations with large size biomolecules. The development framework relies on the collaborative server launchpad.net together with the distributed version control system bazaar.

### *Results*

ATTRACT is the rigid-body docking program based on PTools. It shows good results with protein-protein and protein-DNA complexes as well as 3-body protein docking. We recently studied the RecA nucleofilament formed on single-stranded DNA (ssDNA) and involved in the homologous recombination of prokaryotes. Docking simulations of double-stranded DNA (dsDNA) on RecA indicate that highly curved dsDNA can access the RecA-bound ssDNA while initially retaining its Watson-Crick pairing.

### *Discussion*

The PTools library is freely available under the GNU GPL license, together with detailed documentation and a step-by-step tutorial. In addition to its ease of use, PTools and its docking program ATTRACT are efficient tools to study large and heterogeneous biomolecular complexes. Our next developments aim to embed the partner flexibility during the docking process and to enhance the multicomponent ability.

### *URL*

*http://ptoolsdocking.sourceforge.net/*

### *Presenting Author*

Pierre Poulain (pierre.poulain@univ-paris-diderot.fr)
Univ. Paris Diderot - Paris 7 & Inserm U665

### *Author Affiliations*

(1) DSIMB, Inserm UMR-S 665 et Université Paris Diderot - Paris 7, INTS, 6 rue Alexandre Cabanel, 75015 Paris, France. (2) MTI, Inserm UMR-M 973, Université Paris Diderot-Paris 7, Bâtiment Lamarck, 35 rue Hélène Brion, 75205 Paris Cedex 13, France. (3) LCMBA, UMR-CNRS UNSA 6001, Faculté des Sciences, Université de Nice-Sophia Antipolis, 06108 Nice Cedex 2, France. (4) Physik-Department (T38), Technische Universität München, James-Franck-Str. 1, 85748 Garching, Germany. (5) Laboratoire de Biochimie Théorique - UPR 9080 CNRS, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, F-75005 Paris, France.

### *Acknowledgements*

This work is funding by the Paris Diderot-Paris 7 University; the National Institute for Blood Transfusion (INTS); the Institute for Health and Medical Care (Inserm); the National Center for Scientific Research (CNRS); research grant ANR-07-CIS7-003-01.

## C-13. Development of a structural alignment tool for protein local surfaces

*Minai R (1,*), Horiike T (1)*

Small molecule drugs generally bind to protein surface regions that have geometrical and physicochemical properties suitable for the structure of each molecule. It is therefore possible that nonhomologous proteins with different overall three-dimensional folds could bind an identical small molecule if they have similar surface regions. Since these cause the cross-activity of drugs, to detect similar local surface regions would be valuable in drug discovery. Here, we have developed an alignment tool that can compare the entire local surface of proteins and identify similar regions.

### *Materials and Methods*

This tool uses a feature vector to define the geometrical and physicochemical properties of solvent-accessible atoms of a protein. The physicochemical property is defined 6 types (cation, anion, hydrogen-bond donor, hydrogen-bond acceptor, hydrophobic, and none of these types) according to the PATTY (programmable atom typer) algorithm. The feature vectors are used for the coordinate representation, the sampling optimization (together with geometric hashing), and the similarity scoring of local surface areas. Similar regions are obtained by single linkage clustering of local similar areas.

### *Results*

The proposed method runs on a personal computer with the input of a pair of PDB files and outputs similarity scores and superposed coordinates. It was tested on several protein pairs that have the same function with different folds, for instance, phosphoenolpyruvate carboxykinase (PDB code: 1k3c, SCOP fold: PEP carboxykinase-like) and ABC transporter (1oxu, P-loop containing nucleoside triphosphate hydrolases), and trypsin (1tpo, Trypsin-like serine proteases) and subtilisin (2prk, Subtilisin-like). As a result, the function sites (e.g., the ATP-binding site) were detected as similar regions.

### *Discussion*

This tool generates not only a similarity score but also a statistical score (Z-score) that describes the degree of similarity between the identified similar regions. It can thus be used for the efficient comparison of similar regions. A high Z-score suggests a high probability that the identified regions are important similar regions. In practice, the number of regions having high Z-scores (≥4) obtained in the test calculations is relatively small (approximately ten).

### *URL*

http://d-search.atnifty.com/research.html

### *Presenting Author*

Ryoichi Minai (drminai@ipc.shizuoka.ac.jp)
Graduate School of Science and Technology, Shizuoka University

### *Author Affiliations*

Graduate School of Science and Technology, Shizuoka University

## C-14. Blender Game Engine as a tool to navigate the conformational space of proteins

*Zini M-F (1), Andrei R (1,2), Loni T (1,3), Zoppè M (1,*)*

Proteins and other macromolecules typically exert their function through motion. Glimpse of these movements can be acquired, e.g., by NMR spectroscopy: this information is delivered in pdb files containing many models of the same molecule. We reasoned that if a protein can acquire many conformations, there must be a series of intermediates connecting them. The tools developed for computer graphics (CG) are now powerful enough to be used in complex elaborations such as those involved in molecular motion, which comprise many 1000s atoms. Blender is a CG program open-source and freely available.

### *Materials and Methods*

Blender is a complete package for 3D computer graphics and video games. The program is open source and can be scripted in Python. We prepared scripts to read pdb files, introduce atoms and bonds as specific constraints, record the motion elaborated by the game engine (which also considers collisions), and export all steps of motions as new pdb files. We also use Swiss-PdbViewer and Gromos force filed to evaluate and (eventually) adjust Blender-calculated conformations. All our scripts and programs will be made available after publication, and distributed along with the version 2.5 of Blender

### *Results*

With Calmodulin as case study, we report that using Blender we can generate new conformations that smoothly transit the protein between different models of the NMR file. The intermediate models are calculated in few seconds, and are very quickly adjusted according to Gromos (see Methods) with few minimization cycles. The procedure consists of a reiterative series of steps to find conformations along a single path. By repeating the procedure with different starting points, the navigation map is generated. The resulting series of models is used to observe the protein in motion (v. poster Andrei)

### *Discussion*

The availability of different conformations for the same macromolecule, opens the possibility of studying transition between the various forms. The use of computer graphics tools conbines the advantages of their flexibility and speed with the programmability, so that features can be introduced and handled by the game engine. We humans are especially equipped to grasp concepts and extract infomration through visual exposure, the possibility of animating molecules, and representing them in photorealistic way (see poster Andrei et al) is likely to facilitate comprehension of protein functioning.

### *URL*

http://www.scivis.ifc.cnr.it

### *Presenting Author*

Monica Zoppè (mzoppe@ifc.cnr.it)
Insitute of Clinical Physiology. CNR

### *Author Affiliations*

1 Scientific Visualization Unit, IFC - CNR, Pisa. Italy 2 Lab of Molecular Biology, Scuola Normale Superiore, Pisa. Italy 3 Big Bang Solutions, Navacchio (Pi). Italy

### *Acknowledgements*

work funded by the Regione Toscana (Italy) grant '3d animation' to MZ. RA supported by Scuola Normale Superiore Pisa.

## C-15. Using data mining to identify structural rules in proteins

*Stelle D (2), Scott LPB (1,*), Barioni MC (1)*

Proteins are, among others, the macromolecules that perform all important task in organism. The three-dimensional structure (tertiary structure) of proteins determines their function it is necessary and important for the specific binding of substrates or other ligands. The number of primary structures deposited in databases is growing faster than our ability to solve the tertiary structures using experimental methods. Efficient computational techniques can aid to predict the protein structures and they can help to understand the folding process.Several works have explored this subject.

### *Materials and Methods*

We designed and implemented a local database (DB), consisting around 20.000 protein extracted from the Protein Data Base (PDB). The data base is composed by protein of five folding class. Are stored in the DB information regarding the size and class of each protein, in addition to amino acids (AA) that compose it, together with the physico-chemical properties of each AA. The AA that compose these proteins were organized in windows of different sizes (7, 11, 15, 21) in the DB, each window is associated with its structural motif (this classification was obtained from the DDSP).

### *Results*

We can observe that the experiments with windows with length 7 and 15 obtained a great number of rules with positive lift.It indicated that smaller structural motifs are more easy to be formed. This is a important result that we are investigating with more details in this moment. We need to test this result with a data base bigger with the other class folding. We can observe that there are not rules with lift negative . Thus, our algorithm did not detect sequences (in this data base version) that forbid a structure motif.

### *Discussion*

This is the first step and the first results of our project and software to extract rules to secondary structures of proteins. We observed that data mining is a effcient thechinique to extract this kind of rule in PDB. We obtained three interesting results: the small structural motifs are easier to be formed in proteins of α-helix folding class, there are a bigger number of small sequences that converge to the same structural motif and the algorithm could not detect similar sequences that forbid a specific motif. We are inlcuding transmembrane protein in data base.

### *Presenting Author*

Luis Paulo Barbour Scott (luisp37@gmail.com)
UFABC - CMCC

### *Author Affiliations*

1-CMCC - Universidade Federal do ABC (UFABC), Brasil 2-Universidade Federal do ABC (UFABC), Brasil

### *Acknowledgements*

Fapesp, UFABC

## C-16. Detecting protein domains of biological assemblies using TopDomain-web

*Senn S (1, *), Sippl MJ (1)*

Examining asymmetric units (ASUs) stored in the PDB reveals that a lot of those ASUs don't hold enough information to fully understand the protein structure. A frequent case are missing chains to establish a homomer. In such a case it's important to gather all of the possible structural information to assign domains to the protein. The biological assembly (BA) defined in the PDB file can provide said structural information. A lot of the frequently encountered problems with domains in the protein classifications are based on missing structural information.

### *Materials and Methods*

TopDomain-web is a tool to analyze and decompose different structural entities of proteins. It's an easy to use web application that can handle even biological assemblies up to 6000 residues (99.8 % of all BAs currently defined in the PDB). The application allows instant access to TopMatch-web to perform structural alignments and the Classification of Protein Structures (COPS) to examine quantified relations of the examined protein structure. It is possible to upload files in PDB format.

### *Results*

Examination of the PDB shows that some ASUs do not offer enough structural information to assign all domains of the proteins correctly. For example the ASU of the PDB-file 3g0o is a single chain monomer. The biological assembly is a homodimer which consists of 3 domains with a central domain consisting of parts of both monomeric chains. The central domain can't be assigned using only the ASU. We observe a gain of compactness for the central dimer domain. We also observe that decomposing asymmetric units yields 7.4% more domains per starting entity than decomposing biological assemblies

### *Discussion*

Examining protein structures with TopDomain-web illustrates the fact that using the correct molecule entity, from which to derive a domain decomposition, is important to yield a stable and consistent domain decomposition. A domain decomposition algorithm analyzing the interaction of atoms in a protein structure will interpret missing parts of structures as lack of interaction and deliver inconsequent results. The loss of domains per entity for large scale automated domain decomposition can be explained with rudimentary domains derived from molecules with insufficient structural information.

### *URL*

*http://topdomain.services.came.sbg.ac.at*

### *Presenting Author*

Stefan Senn (steff@came.sbg.ac.at)
Division of Bioinformatics, Department of Molecular Biology, University of Salzburg

### *Author Affiliations*

(1)Division of Bioinformatics, Department of Molecular Biology, University of Salzburg.

## C-17. Metals in protein structures: classification and functional prediction

*Andreini C (1,*), Cavallaro G (1)*

The PDB contains many metal-binding structures reflecting the importance that metals play in proteins. It was estimated that a large fraction of proteins encoded by living systems are metalloproteins, and that about 40% of all structurally characterized enzymes perform metal-dependent reactions. Notwithstanding, bioinformatics resources devoted to the study of metals in biological systems have been scarce so far, most likely due to the difficulty to establish formal criteria to describe the exceptional variety of metalloproteins. Our research is intended to fill this gap.

### *Materials and Methods*

Metal sites present in PDB protein structures are described as 3D templates composed of the metal(s) in the site and the residues forming the first and the second coordination sphere of the metal(s). The sites are grouped based on CATH and SCOP classifications, and the relevant literature is examined on a per-group basis to annotate metal functions and to identify non-physiological metals. Representative sites are selected for each group and compared via structural alignment.

### *Results*

The method was applied to non-heme iron (Fe), copper (Cu) and zinc (Zn) sites. It resulted in the definition of 86 groups of Fe sites, 35 of Cu sites, and 368 of Zn sites. The most common metal function was electron transfer and catalytic for Fe (30% each), catalytic for Cu (ca. 50%), and structural for Zn (55%). The comparison of sites taken from different groups led to cluster together sites which, though present in unrelated proteins, are similar and have the same function. The functionally annotated datasets of metal sites were used to predict the function of structural genomics proteins.

### *Discussion*

The systematic description of metal sites as 3D templates lends itself to perform several analyses that are useful to understand the properties, functions and evolution of metalloproteins. They range from the analysis of metal coordination to the investigation of the evolutionary pathways of metal usage in proteins. Also, structural similarity searches focused on metal sites can be advantageously combined with classical whole-domain comparisons to predict protein function. Finally, this description represents an ideal framework for the systematic classification of metalloproteins in databases.

### *Presenting Author*

Claudia Andreini (andreini@cerm.unifi.it)
University of Florence

### *Author Affiliations*

1. CERM - University of Florence (Italy)

### *Acknowledgements*

Italian Ministry of University and Research (MIUR) through the project FIRB "Futuro in Ricerca"

## C-18. Intuitive visualization of surface properties of moving proteins

*Andrei R (1,2,*), Callieri M (3), Zini M-F (1), Loni T (1,4), Zoppè M (1)*

Surface properties of macromolecules, such as Electrostatic Potential and hydropathy are important features that determine their interaction with the medium or other macromolecules. Proteins are usually in continuous movement, changing shape as they interact with ligands and other molecules, therefore their surface characteristics change in time. A method that permits the visualization of moving molecules simultaneously with their surface properties should facilitate molecular behavior.

### *Materials and Methods*

A harmonious combination of chemico-physical programs, computer graphics software and custom scripts and programs was used for calculating and graphically representing the surface properties of proteins and glycoproteins. Electrostatic Potential and Molecular Lipophilic Potential are visualized on the molecular surface, taking advantage of a free, open-source software, Blender. For their representation we combine computer graphics features such as materials, textures, lights to obtain a photorealistic rendering that give substance and consistency impression.

### *Results*

EP, calculated using APBS, is visualized as animated line particles flowing along field lines from positive to negative, respecting the convention in physics. A novel code is introduced for the MLP visualization: a range of optical features that goes from dark-dull-rough surfaces for the most hydrophilic areas to bright-shiny-smooth surfaces for the most hydrophobic ones. MLP calculation is done using Testa formula based on atomic properties: this permits a spatial distribution of hydropathy more accurate than the well-known Kyte-Doolittle values.

### *Discussion*

EP and MLP are shown simultaneously for moving proteins avoiding the use of color, which cannot be interpreted without a legend. Using real world tactile/sight feelings, the nanoscale world of proteins becomes more understandable, familiar to our everyday life, making it easier to introduce “un-seen” phenomena (concepts) such as hydropathy or charges, while leaving the utilization of color space for the description of other biochemical information. The proteins motion is obtained using the method described in Zini et al.

### *URL*

http://www.scivis.ifc.cnr.it

### *Presenting Author*

Raluca M Andrei (r.andrei@sns.it)
Scuola Normale Superiore

### *Author Affiliations*

1 Scientific Visualization Unit, IFC - CNR, Pisa. Italy 2 Lab of Molecular Biology, Scuola Normale Superiore, Pisa. Italy 3 Visual Computing Lab, ISTI - CNR, Pisa. Italy 4 Big Bang Solutions, Navacchio (Pi). Italy

### *Acknowledgements*

Funded by the Regione Toscana (Italy) grant '3d animation' to MZ. RA supported by Scuola Normale Superiore Pisa.

## C-19. Hierarchical classification of helical distortions related to proline

*Rey J, Devillé J, Chabbert M (*)*

The presence of proline in a helix may be accommodated by several types of distortions. Tools able to predict these distortions would be of great help in molecular modeling. To develop such tools, a systematic classification of proline-related helical distortions is necessary. We have previously shown that helical distortions can be described in terms of structural motifs in which two helices are joined by a few linker residues (Devillé et al., 2008; Rey et al., 2010). This description provides a clue for data mining and can be used to develop a classification method.

### *Materials and Methods*

We analyzed a representative subset of the Protein Data Bank including 4323 protein chains and 26380 helices. Using DSSP, we searched this subset for both contiguous helices and helical HXnH motifs containing proline. Helical HXnH motifs consist of two helices joined by n residues whose dihedral angles were located in the generous alpha conformation. These motifs were classified as a function of the number of linker residues, of their dihedral angles and of the proline position. Analyses were carried out using home-developed scripts.

### *Results*

"Typical" or "non-typical" distortions correspond to proline in contiguous helices and in helical HXnH motifs. They can be differentiated by DSSP. "Non-typical" distortions can be further classified as a function of the dihedral angles of the linker residues. This classification is equivalent to classification based on main chain – main chain H-bonds and differentiates pi bulges and tight turns. Sub-division of bulges and turns is based on the number of linker residues and on the position of proline in the second helix. Using these parameters, about 85% of "non-typical" proline distortions are described by only five canonical structures.

### *Discussion*

We developed a hierarchical method that allows classifying proline distortions into a limited number of canonical structures in three steps. The first step is based on DSSP whose algorithm yields a robust evaluation of H-bonds. The following steps are based on discrete criteria, resulting in an efficient classification scheme. This one can be easily implemented on a large scale and will help to develop predictive tools for molecular modeling.

### *Presenting Author*

Marie Chabbert (marie.chabbert@univ-angers.fr)
CNRS UMR 6214 - INSERM U771

### *Author Affiliations*

CNRS UMR 6214 - INSERM U771

### *Acknowledgements*

We thank NEC Computers Services SARL (Angers, France) for the kind provision of a multiprocessor server.

## C-20. A “smooth” knowledge based potential for protein structure refinement

*Røgen P*(1), Koehl P(2)*

The accuracy of a predicted protein structure is essential for its biological usefulness. It is first within the last few years that positive results on structure refinement have been reported in the literature. Phenomenological, it seems that the local roughness of potential energy functions causes a refinement algorithm to go off in a random direction rather than bringing the model closer to the native structure. To enable the use of gradient based refinement methods the potential is modeled using splines and is smooth except for the salvation term which is only piecewise smooth.

### *Materials and Methods*

Our carbon-alpha knowledge based potential contains 3 different types of terms: A local term, which for each 7-mer in the quarry sequence expresses its structural preferences and flexibility as extracted from a set of known 7-mer structures. The local term mainly expresses the shortest distance to known configurations. A non-local pair potential. And finally solvent effects depending on solvent accessibility and occupied volume. We present a novel method for constructing a knowledge based potential by minimizing the roughness of minima around native structures using 400 models for each of 1326

### *Results*

Comparative modeling techniques can give relatively good models if a homologous protein has known structure. In such cases our potential and the RMSD to the native structure have a correlation of 0.8 to 0.9 which seems promising for structure refinement. This very high correlation stems from a sampling of the local 7-mer geometry of known structures. In the range of free modeling where no prior data is available the average energy-RMSD correlation drops to 0.72 but stays in the range of 0.8-0.9 for most folds.

### *Discussion*

It is well known that a knowledge based potential based on decoy structures have a tendency to learn this specific type of decoys. Hence our method needs testing on other decoy sets but mostly when used for structure refinement. In the absence of prior knowledge the performance often remains good. It’s an interesting problem to determine why and when the performance occasionally is poor. The functional form of the potential makes it highly suited for threading which in an interesting application to pursue.

### *Presenting Author*

Peter Røgen (Peter.Roegen@mat.dtu.dk)
Technical University of Denmark Departmant of Mathematics

### *Author Affiliations*

(1) Technical University of Denmark, Department of Mathematics (2) Department of Computer Science, Genome Center, University of California Davis, USA

## C-21. PROTEIN PEELING in 2010: recent developments and applications

*Gelly J-C (1,*), de Brevern A G (1)*

To better understand the architecture and anatomy of proteins we developed Protein Peeling web server to sub-divide a protein structure in smaller sub-units. These elements called Protein Units (PUs) allow analysis of protein structure organization in simple and detailed description. We propose here a new version of Protein Peeling web server incorporating new progress: unstructured terminal segments recognition, novel scoring function for PUs characterization and lastly structural domains identification. PUs can be used to study protein structure folding, stability or evolution.

### *Materials and Methods*

Based on a new refined non-redundant protein structure databank, we have defined a new assignment method to identify unstructured N or C termini segments: These PUs are identified as unstructured if they are isolated at the first cutting event and no more thereafter. Additionally, a characterization of PU by pseudo-energetic criterion based on statistical potentials computation on carbon alpha is proposed. Lastly a new bottom up algorithm called Domain Reconstruction (DR) provides domain identification. Results have been compared to classical protein domain benchmark datasets.

### *Results*

Protein Peeling method had been implemented in a webserver. It takes a protein structure (PDB format) and gives cutting events of PU using various visual outputs: cutting dendrogram with the different generated PUs, secondary structure contents, visual representation of PUs. An update of the interface that provides tools for viewing and detailed new analyses was developed. Hence, unstructured N or C termini segments, the novel characterization of PU by pseudo-energetic criterion and domain identification are all available. Some practical cases are shown.

### *Discussion*

Protein structures are often described as series of alpha-helices and beta-sheets, or at a higher level as an arrangement of protein domains. We have proposed an intermediate view, the Protein Units (PUs). They are novel level of protein structure description between secondary structures and domains. PUs are linear and compact sub-region of the structure defined by a high number of intra-PU contacts and a low number of inter-PU contacts. Our new developments can help in defining (i) mobile extremities, (ii) pertinent energetic assessment of PUs, and (iii) given a new view of Protein Domains.

### *URL*

*http://www.dsimb.inserm.fr/dsimb_tools/peeling3*

### *Presenting Author*

Jean-Christophe Gelly (jean-christophe.gelly@univ-paris-diderot.fr)
DSIMB, Inserm UMR-S 665 and Université Paris Diderot – Paris 7, INTS, Paris, France

### *Author Affiliations*

DSIMB, Inserm UMR_S 665 and Université Paris Diderot – Paris 7, INTS, Paris, France

### *Acknowledgements*

Financial support was provided by grants from the French Ministry of Research, Paris Diderot - Paris 7 University, the National Institute for Blood Transfusion (INTS), the Institute for Health and Medical Research (INSERM) and Parternariat Hubert Curien (PHC) Orchid.

## C-22. Local moves for efficient sampling of protein conformational space

*Bottaro S (1,*),Boomsma W(1) ,Enøe Johansson K (2) , Andreetta C(2 ), Hamelryck T (2), Ferkinghoff-Borg J (1)*

During their function proteins often undergo local conformational rearrangements. Standard molecular dynamics simulations cannot address the typical time scale at which these processes occur (10^-6 - 10^-3 s). Monte Carlo simulation is an alternative methodology to compute pathways and thermodynamic properties of proteins. To probe the conformational space around the densely packed native state a method to perform local deformations of protein backbone is required

### *Materials and Methods*

We divide the local move into two steps. During the first step new angles are proposed for a small segment of the chain, introducing a break at the end of the segment. In the second step are found the changes in the remaining degrees of freedom of the segment in order to return to a closed state. We found an analytical solution for the second step (loop or chain closure), which is used, in turn, to derive an efficient strategy to ensure the success the chain closure. The move is used in Monte Carlo simulations of proteins using OPLSaa potential in combination with the GB/SA solvation model.

### *Results*

We verify the formal correctness of the method, showing the geometrical validity of the approach. A comparison with the existing state of art methodology is performed , and an improvement of a factor 30 in simulation efficiency was found. A set of Monte Carlo runs on ubiquitin were able to reproduce with statistical significance data obtained from NMR experiments.

### *Discussion*

The results obtained suggest that the move is suitable for establishing a correspondence between MC trials and real molecular time scales, thus providing dynamic information within a MC framework. Although our research was focused on MC simulations, it should be noted that this method may be useful for any application where the resampling of protein conformations is needed, such as loop modeling or structure refinement.

### *URL*

*http://www.phaistos.org/*

### *Presenting Author*

Sandro Bottaro (sbo@elektro.dtu.dk)
Techincal University of Denmark - DTU Elektro

### *Author Affiliations*

(1) DTU Elektro, Technical University of Denmark, 2800 Lyngby, Denmark, and (2) Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark

### *Acknowledgements*

AS Radiometer

## C-23. Self-organising maps applied to protein structure classification

*Alanis-Lobato G (1,2,*)*

Protein Structure Classification is very important for pharmacists, medical doctors, biologists, chemists, and others. Because of this, several databases, methods and algorithms are devoted to the organisation of protein structures according to its architecture, functionality, and other features. These tools normally require a lot of manual input to classify which makes the task tedious. The main motivation of this work is to ease the classification of protein structures by making use of the unsupervised neural network called SOM, which only requires the protein PDB or its amino acid sequence

### *Materials and Methods*

The sample set for the SOM (based on 8192 representative proteins from the SCOP classification) was generated using BioPerl which obtains the amino acid sequence of proteins from PDB files. Once obtained, the sequences were represented with 3 different type of vectors, taking into account amino acid properties such as solubility and volume (linked to the hydrophobic and steric effects respectively), and position in the sequence. The Matlab's SOM Toolbox was used to apply the algorithm and obtain visual results.

### *Results*

Different SOM sizes were used to cluster the 8192 sample proteins. All validations were made by comparing proteins organised in one group by the SOM with the classes proposed by SCOP. Interestingly, no matter the SOM size, around 6 clusters were generated, which pretty much matched the SCOP classes (all alpha, all beta, a/b, a+b, multi-domain, and membrane proteins). There were, of course, some not very large extra clusters whose proteins could represent either some of the other SCOP classes (like coiled coil, peptides, or designed proteins) or totally new categories.

### *Discussion*

The results from this work have shown that SOMs are very powerful tools to cluster proteins with similar structures starting from something as simple as the vector representation of their amino acid sequence. Once trained, SOMs are able to cluster a new protein in a group with similar characteristics. This can be exploited to predict functionality of proteins, given the relation structure-function that holds for lots of them. Since this prediction might be unreliable given the simplicity of the vector representations, SOMs can be part of preprocessing steps in more powerful predicting tools.

### *Presenting Author*

Gregorio Alanis-Lobato (gregorio.alanislobato@kaust.edu.sa)
King Abdullah University of Science and Technology

### *Author Affiliations*

(1)Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia (2) ESCOM - National Polytechnic Institute, Mexico City, Mexico

### *Acknowledgements*

King Abdullah University of Science and Technology

## C-24. Construction of a sequence- and backbone-dependent rotamer library by hidden Markov model

*Wen W(1),Lv Q(1,2,*),Yang P(1),Yang L-Y(1),Wu J-Z(1),Huang X(1)*

However, the state-of-the-art rotamer libraries focus on the statistical information of individual amino acids, ignoring the direct affection of its adjacent amino acids. This article presents a sequence- and backbone-dependent rotamer library. Both the conformation information of adjacent amino acids and torsion angle of the current residue are taken into account to construct a sequence- and backbone-dependent library by HMM.

### *Materials and Methods*

The sequence- and backbone-dependent model considers not only amino acids sequence information but also backbone dihedral angles as observed data. The model is trained to learn the relationships among the observed data and side-chain conformation. After being trained based on comparative training data, this model can produce the specific rotamer library given the target amino acids sequence by sampling the two trained models.

### *Results*

Comparing to rotamer library produced by the other popular side-chain modeling methods, ours is more close to native conformation.We find that the prediction accuracy outperforms on all the test targets to a certain extent comparing with that based on traditional backbone dependent rotamer libraries.

### *Discussion*

This thesis verifies that the sequence- and backbone-dependent model will enable fully taking into account the possible affection of sequence amino acids to the specific amino acid side-chain. The rotamer libraries generated by sequence- and backbone-dependent models provide solid support to protein side-chain prediction.

### *Presenting Author*

Qiang Lv (qiang@suda.edu.cn)
School of Computer Science and Technology, Soochow University

### *Author Affiliations*

1 School of Computer Science and Technology, Soochow University, Suzhou, 215006, China 2 Jiangsu Provincial Key Lab for Computer Information Processing Technology, Suzhou, 215006, China

### *Acknowledgements*

This work is supported by National Natural Science Foundation of China under the grant number 60970055.

## C-25. Codon usage and protein structure in prokaryotes: old stories, new findings and puzzling questions

*Cozzetto D (1,*), Ward S (1), Jones DT (1)*

The community has long discussed whether synonymous codon usage and protein secondary structure are related or not, by supporting or refuting it based on diverse datasets and methodologies. Early work suggested that short segments of more/less common codons strongly correlates with the observation of helices and strands respectively in E. coli highly expressed genes. Though challenging this view, subsequent studies did not clearly disprove it, thus leaving the question still unanswered.

### *Materials and Methods*

We collected 408 E. coli globular single domain protein structures that are confidently and completely mapped to their coding sequences. We used codon frequencies as a proxy for their respective tRNA abundance (and hence translation speeds) and calculated their sums for all tricodons – segments of 3 consecutive codons. We used chi-squared tests to compare the distributions of observed tricodons with those simulated by 250 independent random samples of the codons associated with individual amino acids in our dataset. We partly extended this study to 1024 completely sequenced bacterial genomes.

### *Results*

i) Coding sequences include tricodons that are of much higher frequency than expected by chance; ii) In the set of lowest frequency tricodons,, helices are significantly underrepresented but strands are not overrepresented. Highly expressed genes show no significant bias; iii) Strands are significantly overrepresented in the set of highest frequency tricodons, while helices are significantly underrepresented. Similar results hold for highly expressed genes, too; iv) Analysis of genome-wide data confirms these results and highlights evolutionary conservation of codon usage patterns.

### *Discussion*

Protein secondary structure does not correlate with codon usage, once the structural biases of the amino acids is considered. Other protein-wide properties like contact order, expression level, and overall secondary structure do not explain the variance in codon usage frequencies in and across prokaryotes. Yet, codon frequencies capture an important signal, being conserved at significant mutational distances and in a number of different folds. Additional work is needed to unveil it, and small (10-30 aa) fragment folding kinetics looks promising to explain differences at the fold level.

### *Presenting Author*

Domenico Cozzetto (d.cozzetto@cs.ucl.ac.uk)
Dept. of Computer Science - University College London

### *Author Affiliations*

(1) Department of Computer Science University College London Gower Street, London, WC1E 6BT

### *Acknowledgements*

EU-funded Marie Curie IEF Project ProtDNABindSpec (PIEF-GA-2009-237292); BBRSC

## C-26. Analysis of FTICR mass spectra with H/D exchange of deprotonated dinucleotides

*Claesen J (1,*), Valkenborg D (1,2), Burzykowski T (1)*

Hydrogen/Deuterium exchange (HDX) combined with mass spectrometry can be used to study the conformational speciation of small biomolecules. By monitoring the level of deuterium incorporation over time, information about the accessibility of hydrogens for the exchange in various parts of the molecule is gathered. Determination of the number of exchanged hydrogen atoms and their exchange rates allows drawing conclusions about the conformation and dynamics of proteins.

### *Materials and Methods*

We propose the following method to estimate the exchange rates (l_1,..,l_n), of n hydrogens: 1)Read mass spectra for all time points in 2)For spectrum k, determine the intensities of observed peaks (O_k,1,...,O_k,n+p) 3)Determine the isotopic distribution of molecule M before HDX has taken place (F_0) 4)For spectrum k, calculate the expected intensities by means of summing the "shifts" of F_0, which occur with probabilities modeled as a function of time t and exchange rates l_i 5) Estimate l's by minimizing the Residual Sum of Squares for the observed and expected intensities over all spectra.

### *Results*

The proposed model and algorithm are evaluated on dinucleotides dTT, dTG and dAA. -dTT: 4 protons are potentially available for HDX; according to our model, 2 exchanged. -dTG: 6 protons are potentially available for HDX; according to our model, all 6 exchanged, but one exchanged at a very slow rate. -dAA: 6 protons are potentially available for HDX; all exchanged at the same rate. These results are in accordance with the results reported by Balbeur et al 2007. However, not all observed intensities can be explained by our model. This might be due to the fact that some protons exchanged before

### *Discussion*

Our method allows predicting the number of HDXs of H atoms and estimating exchange rates for the available protons of a dinucleotide or protein.

### *Presenting Author*

Jürgen Claesen (jurgen.claesen@uhasselt.be)
Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt

### *Author Affiliations*

(1) Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium (2) VITO

## C-27. Voroprot: an interactive tool for the analysis of geometric features of protein structure

*Olechnovič K (1, *), Margelevičius M (1), Venclovas Č (1)*

The ability to analyze and visualize geometric features of the three-dimensional protein structure is fundamental for studies of protein folding, residue packing, docking, protein-protein interactions, etc. Some of the advanced geometric data structures, like Apollonius diagram (also known as additively weighted Voronoi diagram), make it possible to perform such an analysis in a comprehensive manner. With this goal in mind, we have set out to develop an interactive protein structure analysis and visualization tool.

### *Materials and Methods*

In Voroprot we used our own algorithm implementation to construct and display Apollonius diagram and Apollonius graph of protein atoms, considering each atom to be a 3D ball of Van der Waals radius. We subdivide a protein structure into quadruples of atoms, making sure that no tangent sphere of any quadruple overlaps any atom. This subdivision is used to find internal cavities and to construct Voronoi cells and various surfaces for each atom. Voroprot is written in C++ using Qt library and OpenGL API.

### *Results*

We developed Voroprot, an easy-to-use cross-platform software tool, that provides a unique set of capabilities for exploring geometric features of protein structure. Our tool concentrates on geometric methods not included in other software packages used for protein structure analysis. Using Apollonius diagram, Voroprot allows construction and visualization of protein alpha shapes, atoms contact surfaces, solvent accessible surfaces and cavities inside protein structure.

### *Discussion*

The use of additively weighted Voronoi diagram offers a powerful approach in studies of protein structure. However, this geometric data structure so far has been exploited by software developers only to a limited extent. Voroprot is a significant advancement in effective utilization of Apollonius diagram for protein structure analysis and visualization.

### *URL*

http://www.ibt.lt/en/laboratories/bioinfo_en/software/voroprot-2.html

### *Presenting Author*

Kliment Olechnovič (kliment@ibt.lt)
Institute of Biotechnology

### *Author Affiliations*

(1) Institute of Biotechnology, Vilnius, Lithuania

### *Acknowledgements*

This work was supported by the Howard Hughes Medical Institute grant to ČV.

## C-28. BBP - Beta-Barrel Predictor: a web server for the prediction of the super-secondary structure of transmembrane β-barrel proteins

*Tran V D (*), Chassignet P, Steyaert J-M*

The transmembrane β-barrel (TMB) proteins perform diverse important functions in cells. Computational structure prediction methods based on learning are poorly tractable for these proteins since they are difficult to observe by standard experimental techniques. Long range structure is also a challenge. Generally, those structures are not only a series of β-strands where each is bonded to the preceding and following ones in the primary sequence, but they may contain Greek key or Jelly roll motifs as well. This may be described as a permutation on the order of the bonded segments.

### *Materials and Methods*

We model the protein folding problem with minimum energy into the search of the longest closed path in a weighted graph with respect to some given permutation. A trained probabilistic model has been introduced and acts as a primary filter to identify substrings as potential β-strands (vertices) or coils (edges). The weights are related to energy terms of electrostatic, hydrophobic and hydrogen bonding interactions. By dynamic programming, the algorithm runs in O(N^2) for the identity permutation, and in at most O(N^4) for the Greek key motifs, where N is the number of amino acids.

### *Results*

We develop a web-based application which is capable to predict the super- secondary structure of transmembrane β-barrel proteins with or without permuted structure. The input is a sequence, a pattern for the strand per- mutation and a training set for the short range statistics. The output also gives the side-chain orientation of β-strand residues, the inter-strand residue contacts and the folding pseudo-energy. Thus, most of α-bundles (97%) and globular β-barrels (81%) are easily rejected and BBP may be used as an effective tool for genome screening.

### *Discussion*

We have implemented a fast pseudo-energy minimization method for trans- membrane protein super-secondary structure prediction based on a variety of potential structures. The model has been tested with TMB proteins and it performs with an efficiency comparable to those of previous approaches. Yet our approach does not rely only on the known structure families; it is more flexible and allows the modeling and exploration of a large variety of structures. We are validating the results on the well known yeast genome and further results are expected for other genomes.

### *URL*

*http://www.lix.polytechnique.fr/Labo/Van-Du.Tran/bbp*

### *Presenting Author*

Van Du T. Tran (vandu@lix.polytechnique.fr)
Laboratoire d'Informatique de l'Ecole Polytechnique

### *Author Affiliations*

Laboratoire d'Informatique de l'Ecole Polytechnique

### *Acknowledgements*

Ecole Polytechnique

## C-29. Prediction of three dimensional structure of protein complexes

*Plewczyński D (1),Łaźniewski M (1,2*), Augustyniak R (2), Ginalski K (1)*

Docking is one of the most commonly used techniques in drug design. It is employed for both identifying correct poses of a ligand in the binding site of a protein as well as for the estimation of the protein-ligand interaction strength. Since millions of compounds must be screened, before identifying suitable target for biological testing, all calculations should be done in a reasonable time frame. Since almost every year new software is released it should be evaluated by the independent group, to confirm advantage of new version. It is also crucial to compare all programs on same data set.

### *Materials and Methods*

Seven popular docking programs (Surflex, LigandFit, Glide, GOLD, FlexX, eHiTS and AutoDock) were tested on the extensive dataset composed of 1300 protein-ligands complexes from PDBbind 2007 database, where experimentally measured binding affinity values are also available. We compared independently the ability of proper posing (according to RMSD of predicted conformations versus the corresponding native one) and scoring (by calculating correlation between docking score and ligand binding strength). To our knowledge it is the first large-scale docking evaluation.

### *Results*

Based on our results we can order tested programs in the following way: GOLD ~ eHITS > Surflex > Glide > LigandFit > FlexX ~ AutoDock. The best programs have the average RMSDtop score around 2.7Å, and it rises to nearly 4.5Å for the weakest FlexX. Percentage of successfully docked pairs, i.e. complexes for which top scored conformation had RMSD below 2Å, is around 60% for best program GOLD, while for weakest FlexX it drops below 40%. As for scoring capabilities Pearson correlation between docking score and ligands true binding affinities is 0.36 for best in this test eHiTS scoring function.

### *Discussion*

Our results clearly show that using existing software ligand binding conformation could be identified in most cases. Moreover programs produce similar docking solutions regardless of ligands starting conformations used for docking. Still we observe the lack of universal scoring function for all types of molecules and protein families. Most promising function in our test eHiTS_score is classified to knowledge-based type of functions and it clearly outperforms functions representing different approaches in scoring like force-field based GoldScore or empirically based ChemScore.

### *Presenting Author*

Michał Łaźniewski (michall@icm.edu.pl)
University of Warsaw

### *Author Affiliations*

1. 1Interdisciplinary Centre for Mathemtaical and Computetiobal Modelling, University of Warsaw, Pawińskiego 5a Street, 02-106 Warsaw, Poland 2. 2Department of Physical Chemistry, Faculty of Pharmacy, Medical University of Warsaw, Banacha 1 Street, 02-097 Warsaw, Poland

### *Acknowledgements*

Calculations were performed at the Interdisciplinary Center for Mathematical and Computational Modelling. This work was supported by Polish Ministry of Science and Higher Education N301 159735 grant

## C-30. Investigating differences in the structural environment of parallel and anti-parallel beta sheets

*Bawono P (1,*), Abeln S (1)*

Beta sheet is the second most common form of regular secondary structure in proteins. Despite numerous studies into the structural specifics of beta sheets, the stabilising factors of parallel versus anti parallel beta sheets are not yet well understood. Further understanding of the stability of beta sheets is important in the analysis of amyloid formation. This project investigates the differences in the environmental structure of parallel and anti parallel beta sheets, and whether these differences account for the stability and relative abundance of parallel and anti parallel beta sheets.

### *Materials and Methods*

In this research we focus on two major structural features: the solvent accessibility and hydrogen bonds. We use DSSP to calculate the solvent exposed surface of each residue. A residue is considered buried if its solvent exposed surface is less than 7 %. We categorize beta sheets into local and non local based on the sequence distance between two adjacent strands, in order to see whether the locality can explain the difference in solvent accessibility. If the distance is greater than 15 residues then the beta sheet is non local and it is local if otherwise.

### *Results*

We found that there is a significant difference in solvent accessibility. There are more buried residues in parallel beta sheets than in anti parallel beta sheets. Typically, anti parallel sheets have half of the residues buried and the other half exposed. We also found that almost all parallel sheets are non local while anti parallel are more local. The difference in solvent accessibility remains when we account for the difference in locality between parallel and anti-parallel sheets. We are currently analyzing the hydrogen bonding patterns for further environmental analysis.

### *Discussion*

We show that there is an unexplained difference in the structural environment of parallel and anti-parallel beta sheets, with anti-parallel beta sheets being more exposed to the solvent. It is well known that anti-parallel strands are generally more local than parallel strands. However, comparisons between parallel sheets with both local and non local anti parallel sheets still show significant difference, leaving the environmental difference between parallel and anti-parallel sheets still unexplained.

### *Presenting Author*

Punto Bawono (pbo300@few.vu.nl)
Vrije Universiteit, the Netherlands

### *Author Affiliations*

1 Centre for Integrative Bioinformatics Vrije Universiteit Amsterdam

## C-31. Protein model quality assessment based on structural and functional similarities

*Konopka BM (1,*), Nebel J-C (2), Kotulska M (1)*

Experimental determination of 3D structures is time-consuming, expensive and has technical limitations. Consequently, computational prediction of 3D structures is a very attractive alternative. If homologue structures are available, computer-based models can be very accurate. Otherwise, threading and ab initio methods can be applied to approximate the structure of the target. Since these approaches generally produce several models per target, the ability to evaluate their quality is essential. Thus quality assessment procedures should be developed alongside new modelling techniques.

### *Materials and Methods*

Since there is a strong structure-function relationship in proteins, knowledge about function should allow discriminating against poor structure candidates. First, the Distance matrix ALIgnment method is used to find molecules that are structurally similar to the investigated model. Next the description of protein functions from Gene Ontology is used to evaluate the functional similarity of the model's structural neighbours and the prediction target. The gathered information is processed in a modified Receiver Operating Characteristic curve framework to estimate the global quality of the model

### *Results*

The application was validated using models submitted to CASP8 (The Critical Assessment of protein Structure Predictions) contest. The observed quality of models was measured with GDT_TS, which is one of the standard metrics used by CASP assessors. Out of 127 released targets, 75 were found to have Gene Ontology annotations, thus these were used in the validation process. Correlation coefficients of proposed measures and GDT_TS, pooled for all targets and for each target separately were calculated: best Rpooled=0.511, best Raverage=0.521.

### *Discussion*

The method performed well discriminating between good and bad models. However, its performance in the task of ranking high quality models was less satisfactory. Since the approach is novel and benefits from sources of information that are not explored by other Model Quality Assessment Programmes, there is a chance that in conjunction with other methods it can significantly increase the accuracy of estimating the quality of protein model-structures. The method can be efficiently employed when neither experimental structure of a protein nor its homologues are known.

### *Presenting Author*

Bogumil M. Konopka (bogumil.konopka@pwr.wroc.pl)
Wroclaw University Of Technology, Faculty of Fundamental Problems of Technology, Institute of Biomedical Engineering and Instrumentation

### *Author Affiliations*

1 Institute of Biomedical Engineering and Instrumentation, Faculty of Fundamental Problems of Technology Wroclaw University Of Technology, Wroclaw, Poland 2 Faculty of Computing, Information Systems and Mathematics, Kingston University, Kingston Upon Themes, United Kingdom

### *Acknowledgements*

The project was supported by the Department of Student Affairs at Wroclaw University of Technology (grant no. 273/2010) and Wroclaw Centre for Networking and Supercomputing (computing grant. no 98)

## C-32. Data driven structure prediction: calculating accurate small angle X-ray scattering curves from coarse-grained protein models

*Stovgaard K (1,*), Andreetta C (1), Ferkinghoff-Borg J (2), Hamelryck T (1)*

Genome sequencing projects are currently expanding the gap between the amount of known protein sequences and structures. The limitations of current high resolution structure determination methods, like X-ray crystallography and NMR spectroscopy, make it unlikely that this gap will disappear in the near future. Small angle X-ray scattering (SAXS) is an established method for routinely determining the structure of proteins in solution, but due to the lower structural information content in this data, additional model constraints are needed for the structural interpretation.

### *Materials and Methods*

In order to calculate the scattering curve for a coarse-grained protein structure, we used the well-known Debye formula. Scattering form factors for each amino acid were needed in this description for computational efficiency. The Top500 set of protein structures was used for training. A maximum-likelihood based MCMC method implemented in the Muninn program was applied to obtain a set of samples from the posterior form factor distributions. TorusDBN, a probabilistic model of protein structure formulated as a dynamic Bayesian network was used to calculate the likelihood of the decoy structures.

### *Results*

Two coarse-grained representations of protein structures were investigated; using two scattering bodies per residue gave significantly better results than using a single body. Obtained form factor point estimates were evaluated for a set of diverse proteins, with resulting SAXS curves on par with the current state-of-the-art program CRYSOL. Furthermore, our method was validated in a Z-score comparison for a set of native-like protein decoys from the TASSER program, showing excellent results. Including prior information from the probabilistic model TorusDBN further improved decoy recognition.

### *Discussion*

We have demonstrated that it is possible to obtain accurate SAXS curves from coarse-grained protein structures and matching estimated form factors without the use of ad hoc correction factors. SAXS curve calculations and decoy recognition experiments show that our method performs on par with the current state-of-the-art program CRYSOL, which requires full-atomic detail. The presented method shows great promise for use in statistical inference of protein structures from SAXS data. Such an approach is currently being implemented in the PHAISTOS software package.

### *URL*

*http://www.phaistos.org*

### *Presenting Author*

Kasper Stovgaard (stovgaard@binf.ku.dk)
The Bioinformatics Centre, University of Copenhagen

### *Author Affiliations*

(1) The Bioinformatics Centre, Department of Biology, University of Copenhagen, Denmark (2) Department of Electrical Engineering, Technical University of Denmark, Lyngby, Denmark

### *Acknowledgements*

The Danish Council for Strategic Research (Program Commission on Nanoscience, Biotechnology and IT; project: Simulating proteins on a millisecond time scale, 2106-06-0009) and the Danish Council for Independent Research (Danish Research Council for Technology and Production Sciences; project: Data driven protein structure prediction, 274-06-0380)

## C-33. OpenStructure: A flexible software framework for computational structural biology

*Biasini M (1,2,*), Mariani V (1,2), Haas J (1,2), Scheuber S (1,2), Schenk A-D (3), Schwede T (1,2), Philippsen A (1)*

We introduce OpenStructure, a flexible software framework for computational structural biology, which emerged out of the need for a solid, yet flexible and versatile toolkit for method development. It combines a C++ based library of commonly with a Python layer and powerful visualization tools. OpenStructure is also designed to easily accommodate interfaces to already existing software. This allows for rapid visually-enhanced prototyping of new functionality, making OpenStructure an ideal environment for the development of next-generation structural biology algorithms.

***Materials and Methods***

The functionality is grouped into modules. Each of these modules consists of a computational core in the form of a shared library of C++ code, a dynamic module that exposes the C++ API to Python using Boost.Python and a set of Python modules built on top of the exported API. Parts of the computational core and the GUI of the Image Processing Library and Toolkit IPLT have been incorporated into the OpenStructure geom, img and gui modules to offer versatile handling of image data with support for various algorithms. Other modules have been written anew to manipulate sequence and structural data.

***Results***

Typically, users interact with a high-level Python interface, while power users with high computational requirements access the API directly at the level of C++ (optionally exporting their algorithms to Python). The framework additionally features capabilities for real-time rendering of molecules, density maps and molecular surfaces which can be visualized in a graphical user interface.

***Discussion***

The flexibility and expressive API of OpenStructure make it an ideal basis for the development of computational methods. For example, new versions of the QMean tools for model quality assessment are based on OpenStructure, as well as the structural variation analysis tools in ProteinModelPortal. Further, work is on the way to implement the next generation of the SWISS-MODEL pipeline using the OpenStructure framework.

***URL***

http://www.openstructure.org

***Presenting Author***

Marco Biasini (marco.biasini@unibas.ch)
Biozentrum, University of Basel

***Author Affiliations***

1 Biozentrum, Universität Basel, Basel, Switzerland. 2 SIB Swiss Institute of Bioinformatics, Basel, Switzerland. 3 Department of Cell Biology, Harvard Medical School, Boston, MA 02115 , USA

## C-34. Modeling structure of ionic channels from rigid network of contact sites

*Kotulska M (1,2, *), Gerstein M (2)*

The ratio of experimental structures of protein channels to their total number is disastrously small compared to other proteins. Lipid environment of the channels and their great size impede the development. Improvement could contribute to drug design and resolving mechanisms of channelopathies. Modeling 3D structure of transmembrane channels will be enhanced by constraints defined by their contact sites distribution. Representing protein as a network of contact sites and characterizing topology of such a channel graph could be employed for limiting protein decoys in modeling process.

### *Materials and Methods*

Protein is represented as a network of non-bond contact sites. As a contact site we consider residues within a defined distance between protein C alpha atoms that are separated by a certain number of amino acids. We test how definition of the contact site, in terms of the distance and separation threshold values, affects topology of the network and which graph model characterizes the protein channel best. The tests are provided for helical channels and barrels. Transmembrane and soluble parts are compared.

### *Results*

Compared to random (Erdős–Rényi) network, graph of a protein shows high clustering coefficient. The average shortest path length is small but greater than in the random model, as in “small-world” graphs, which describe all real world networks. The node degree in protein graphs changes from Poisson distribution ( “small-world” ) to power-law (random graphs) as the residue separation increases. Graphs with high value of the distance threshold also tend from power-law to Poisson distribution. Transmembrane helices contribute power-law, while other components Poisson distribution.

### *Discussion*

The graph model of ionic channel shows distinct features, related also to its secondary structure and location. This could be utilized in protein structure modeling process with regard to contact site distribution. The network topology proved sensitive to the contact site definition, showing that protein graph obeys typical real world network characteristics only if the distance is below certain threshold and the separation above a limit number, and follows random network characteristics otherwise. The thresholds should be incorporated into optimum definition of a contact site.

### *Presenting Author*

Malgorzata Kotulska (malgorzata.kotulska@pwr.wroc.pl)
Wroclaw University Of Technology

### *Author Affiliations*

(1) Institute of Biomedical Engineering and Instrumentation, Wroclaw University of Technology, Wroclaw, Poland; (2) Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, USA

### *Acknowledgements*

MK would like to acknowledge the Fulbright scholarship to Yale University .

## C-35. Multi-task sequence labeling for protein annotation

*Maes F, Becker J (*), Wehenkel L*

Many tasks related to protein annotation can be formalized as sequence labeling problems where the input is a sequence of amino acids and the output is a corresponding sequence of labels (secondary structure, solvent accessibility, disordered region, etc.). To solve $k$ such tasks with machine learning techniques, we can independently learn $k$ sequence labeling models. Several textit{multi-task} approaches that {em jointly} learn multiple related problems have been proposed in the machine learning community recently and have been shown to often outperform single-task approaches. We propose one such model for protein annotation.

### *Materials and Methods*

We introduce an original multi-task approach for sequence labeling problems that iteratively re-estimates each target sequence several times, using the input and the current state of all other predicted sequences. * Each label is estimated based on a surrounding window of input elements and of predicted elements in other target sequences. The label-predictor is a linear support vector machine, learned with stochastic gradient descent. * Our study relies on two datasets. The first one includes 500 proteins extracted from PDB (at 70% identity). The second one is the CB513 dataset which enables to compare our results with the state-of-the-art of secondary structure.

### *Results*

To illustrate our novel approach, we focus on the joint prediction of the secondary structure (with both three and eight classes), the solvent accessibility (with two classes) and the disorder regions (with two classes). We show experimentally that the multi-task approach nearly always outperforms single-task sequence labeling on the PDB dataset.

### *Discussion*

We show that solving simultaneously multiple protein annotation problems leads to an improvement over methods that treat each problem independently. When combining two tasks, the results are nearly always better than corresponding single-task models. However, when treating more than two tasks together, we often observe a slight degradation of results. This might be due to an over-fitting problem. Our future work includes making experiments on larger sets of proteins.

### *Presenting Author*

Julien Becker (J.Becker@ulg.ac.be)
Systems and Modeling - Department of EE and CS, University of Liege, Belgium

### *Author Affiliations*

Systems and Modeling - Department of EE and CS, Systems Biology and Chemical Biology - GIGA-Research, University of Liege, Belgium

### *Acknowledgements*

This paper presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modelling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State and of the PASCAL network. Julien Becker is recipient of a F.R.I.A. fellowship.

## C-36. A convex programming model for protein structure prediction

*Kauffman C (*), Karypis G*

Most current protein structure prediction methods involve high computational cost, mostly due to the nonconvex function used to represent the protein's potential energy surface. Though thought to be chemically accurate, such functions result in may local minima which must be searched to find the global minimum energy conformation, generally regarded as where the native state of the protein will reside. We propose an alternative in which we can guarantee protein structures are at the global minimum and find them quickly, the cost being interpretability of the energy function.

### *Materials and Methods*

Our approach is to represent protein structure conformations in a semidefinite program (SDP), a convex optimization problem in which a global minimum can be found efficiently. Bond lengths become linear constraints in the program and atom clashes are avoided using inequality constraints. The energy of the conformation is represented as a weighted sum of the squared distances between unbonded atoms. Inverse optimization may be used to efficiently determine energy function parameters which place a structure at the energy minimum.

### *Results*

One application of our convex framework is to recover coordinates from distance information on structures, potentially a useful capability for NMR experiments. In an evaluation study on 124 known structures, our method produced statistically lower RMSDs than two standard distance geometry techniques, though our method needs to be adapted to noisy and sparse distance matrices which are usually the case for NMR experiments. We are also conducting a study of sequence-based structure prediction on a set of 2748 proteins using convex optimization.

### *Discussion*

Our work is an attempt to determine how well protein structure can be estimated using convex optimization techniques. Restricting ourselves to the class of convex functions ensures that at the very least the resulting technique will be fast enough for most desktop workstations to produce structure predictions in a matter of minutes to hours. Our preliminary results indicate that there is potential for convex optimization to produce accuracy that is competitive with standard homology modeling and ab initio structure prediction techniques.

### *URL*

*http://bioinfo.cs.umn.edu/supplements/eccb2010*

### *Presenting Author*

Chris Kauffman (kauffman@cs.umn.edu)
University of Minnesota, Twin Cities

### *Author Affiliations*

University of Minnesota, Twin Cities

### *Acknowledgements*

National Institute of Health [T32GM008347, RLM008713A], National Science Foundation [IIS-0431135], and the University of Minnesota Digital Technology Center

## C-37. Predicting protein function with the relative backbone position kernel

*Schietgat L (*), Aryal S, Ramon J*

Proteins are macromolecules that play crucial roles in many biological processes. As more data about proteins become available, the automatic classification of their function is an important challenge in bioinformatics. A lot of techniques have been proposed that predict function based on the primary and secondary structure of the proteins. However, little attention has gone to 3D structures of proteins, while these carry a lot of additional information. Recently, it was shown that kernel methods provide good results on protein classification tasks. However, none of them use 3D information.

### *Materials and Methods*

We propose a new kernel for proteins called the relative backbone position kernel (RBPK). It makes use of 3D information by comparing Euclidean distances between the residue atoms and the backbone atoms of the protein. In this way, the kernel can select spatial features that are important for interactions with ligands or other proteins and will influence protein function. We evaluate our kernel through two datasets: one contains protein structures that have to be classified into enzymes and non-enzymes, while for the second one, the task is to predict the resistance of HIV protease structures.

### *Results*

We compare the performance of RBPK with the Fast Subtree Kernel (FSTK), which is a state-of-the-art kernel for protein function classification. FSTK uses a graph with amino acids as vertices and distances as edges for the representation of the proteins. Although FSTK is more efficient to compute than RBPK, the latter obtains a higher predictive accuracy, resulting in state-of-the-art results for the two datasets.

### *Discussion*

Our experimental results show that RBPK, wich exploits 3D information of the protein, leads to more accurate predictions over a recently proposed graph-based kernel. The accuracy of above 85% for the first dataset (D&D) is an encouraging result. However, the computational efficiency of the kernel still remains an important issue, especially for large proteins. There are several ideas to improve this, by limiting the amount of distances that are computed by the kernel.

### *Presenting Author*

Leander Schietgat (leander.schietgat@cs.kuleuven.be)
Katholieke Universiteit Leuven

### *Author Affiliations*

Department of Computer Science, Katholieke Universiteit Leuven

### *Acknowledgements*

This research is funded by the ERC Starting Grant 240186: Mining Graphs and Networks: a Theory-based approach.

## C-38. The Torsional Network Model: improvements in modelling of protein flexibility

*Méndez Giráldez R (*), Bastolla U*

Proteins undergo significant conformational rearrangements to achieve their function, which often pose serious limitations to protein structure prediction and docking methods. Those collective motions have been assessed through normal mode computations, using typical all atom force fields. Elastic Network Models in Cartesian space simplify computations but still yield reasonable results, at expense of distorting protein secondary structure. Our TNM increases protein structure accuracy while reducing the number of degrees of freedom, by only allowing main backbone torsion angles to change.

### *Materials and Methods*

The matrices of the second derivatives of the Elastic Network potential and kinetic energy with respect to torsion angles are calculated as linear transformations of the corresponding matrices in Cartesian space through the Jacobian. The latter is restricted to internal degrees of freedom imposing Eckart conditions. Normal modes in torsional space are obtained by solving a generalized eigenvalue problem. TNM modes are compared to ANM ones in terms of: tolerance to elongation, their ability to account for observed conformational changes and the energy barrier reduction in such processes.

### *Results*

TNM modes are in average more robust with than ANM ones, as witnessed by Lennard Jones energy calculations, in a set of 500 protein structures. Low frequency modes are more deformable than high frequency ones, account better for conformational changes in a different set of 50 protein structures and are localized in torsional space, but collective in Cartesian space. Transport and enzyme proteins seem to reduce significantly their conformational change barriers as a result of the selective pressure. Their collective motions can be described by few modes (fewer for the TNM than the ANM).

### *Discussion*

The TNM allows better protein motion characterization than conventional Cartesian Elastic Network Models, and reduces its computational cost. TNM modes (i.e. low frequency ones) localized in torsional space and collective in Cartesian space may be reminiscent of hinge motions. In general, a smaller number of (low frequency) TNM modes than ANMs, is needed to describe the same conformational change. Evolution seems to reduce energy barriers via few low frequency modes that contribute significantly to the change. The results make TNM ideal to model protein flexibility in Computational Biology.

### *Presenting Author*

Raúl Méndez Giráldez (raulmendez@cbm.uam.es)
Centre for Molecular Biology "Severo Ochoa"

### *Author Affiliations*

Bioinformatics Unit Centre for Molecular Biology "Severo Ochoa" Universidad Autónoma de Madrid C/ Nicolás Cabrera, 1 Cantoblanco 28049, Madrid SPAIN

### *Acknowledgements*

Centrosome 3D (Spanish Ministry of Science), CSIC (Spanish Supreme Council for Scientific Research) through the JAE program

## C-39. MEDELLER: coordinate generation for membrane proteins

*Kelm S (1,*), Shi J (2), Deane CM (1)*

Membrane proteins are important drug targets but knowledge of their exact structure is limited to relatively few examples. Existing homology-based structure prediction methods are designed for globular, water-soluble proteins. However, we are now beginning to have enough membrane protein structures to justify the development of a homology-based approach specifically for them. We present a homology-based coordinate generation method, MEDELLER, which uses membrane protein-specific information, and compare it to the popular software Modeller.

### *Materials and Methods*

Comparison between the two methods was performed on a set of over 1000 target-template pairs of membrane proteins, which were classified into four test sets by their sequence identity. MEDELLER first builds a core model, using membrane protein-specific information. Then, loops are added using the FREAD loop structure prediction algorithm, to generate high-accuracy models.

### *Results*

MEDELLER's accuracy distribution over all test sets is clearly shifted towards the more accurate end, when compared to Modeller, with an average backbone RMSD of 2.61Å versus 3.02Å for Modeller. On our “easy” test set, MEDELLER achieves an average accuracy of 0.89Å backbone RMSD, versus 1.55Å for Modeller.

### *Discussion*

The MEDELLER algorithm is conceptually simple and outperforms the popular structure prediction method Modeller on all test sets. Its major advantage lies in the generation of an accurate core model. Future versions of MEDELLER will also address the problem of helix kinks and twists, promising a better overall model quality in cases where the template differs slightly from the target structure.

### *URL*

*http://imembrane.info*

### *Presenting Author*

Sebastian Kelm (kelm@stats.ox.ac.uk)
University of Oxford

### *Author Affiliations*

(1) Department of Statistics, University of Oxford, Oxford, UK (2) UCB Celltech, Slough, UK

### *Acknowledgements*

Biotechnology and Biological Sciences Research Council

## C-40. Structure-based predictor of HIV coreceptor tropism

*Bozek K (1,*), Lengauer T (1), Domingues FS (1)*

Human immunodeficiency virus (HIV) cell entry is the first step of infection and conditions further successful replication. In addition to the main cellular receptor CD4 the virus attaches to one of the coreceptors CCR5 or CXCR4. Viruses binding to CCR5 have been shown to be predominantly present during the early asymptomatic stage of infection. CXCR4-binding viruses are associated with disease progression. New HIV treatment strategies block CCR5 coreceptor. Monitoring of the virus coreceptor usage and understanding the mechanisms driving coreceptor switch are therefore of crucial importance.

### *Materials and Methods*

We built a structural descriptor of the V3 loop, the main determinant of viral tropism, to predict coreceptor usage and to gain more insight into its molecular background. The amino acid properties of V3 loops were mapped on the published reference crystal structures. Each V3 loop was described as a concatenated vector of amino acid properties in its structural parts. We next performed feature selection to reduce the redundancy of highly correlated properties and to extract biologically meaningful structural and physicochemical determinants of coreceptor tropism.

### *Results*

This systematic search for structural determinants of the HIV tropism resulted in an improved prediction method allowing for biological interpretation. The initial prediction based on the full set of descriptors had a similar accuracy to the sequence-based method with lower efficiency due to the long vector representation of each of the sequence. Reducing the feature set increased the efficiency of the method and improved prediction accuracy both in the clonal and clinical datasets. The selected features point to structural parts of the loop and its properties crucial for the coreceptor usage.

### *Discussion*

Structural descriptors are a more accurate measure of the actual physicochemical similarities among amino acids than the 20 letter representation used in the sequence-based methods. The proposed coreceptor prediction method operates on a relatively simple and efficient algorithm not involving structure modeling. The features selected out of the full set of structural descriptors distributed along the V3 loop, not only improved the computational efficiency of the prediction procedure but also indicated which of the biochemical properties on which part of the loop are crucial for the tropism.

### *Presenting Author*

Katarzyna Bozek (bozek@mpi-inf.mpg.de)
Max Planck Institute for Informatics

### *Author Affiliations*

(1) Max Planck Institute for Informatics, Computational Biology and Applied Algorithmics

## AUTHOR INDEX