ECCB 2014 Accepted Posters with Abstracts

E: Structural Bioinformatics

E01: Ruben Acuna, Zoe Lacroix and Jacques Chomilier. SPROUTS 2.0: a database and workflow to predict protein stability upon point mutation

Abstract: Amino acid substitution is now considered as a major constraint on protein evolvability, while it was previously admitted that most positions can tolerate drastic sequence changes, provided the fold is conserved. Actually, mutations affect stability and stability affects evolution. The level of deleterious mutations can be as high as one third. Therefore, the prediction of the effects of residue substitution can be of great help in wet labs and we placed our efforts in proposing an enhanced version of a database devoted to this purpose. In this paper, we focus only on the thermodynamic contribution to stability, which can be considered as acceptable for small proteins. The SPROUTS database compiles the predictions from various sources calculating the DeltaDeltaG due to a point mutation, together with a consensus from eight distinct algorithms.
Due to evolution, the number of stabilizing mutations is smaller than for destabilizing ones. One must mention that a stabilizing mutation is not necessarily related to an improved efficiency of the mutated protein, as far as function is concerned. Sometimes, a more stable structure results in an increased rigidity, while the function requires a certain level of flexibility. This is the case for instance with the enzyme catalysis. Therefore, it seems reasonable to place a threshold of 2 kcal/mol in either ways of DeltaDeltaG (stabilizing or destabilizing) in order to claim to a putative malfunction. Mutations in conserved positions usually cause large stability decreases. This is the reason why we introduced in SPROUTS a supplementary "point of view" from a simulation designated at predicting the importance of a given position for the contribution to the folding process.
The first SPROUTS database compiled in 2008 presented data that capture representative folds and results related to the prediction of critical residues expected to belong to the folding nucleus of 429 structures produced by 7 programs. SPROUTS 2.0 has grown to over 1,300 structures and is queried daily by users from over 58 countries. The new workflow and database have been updated with the most recent versions of the initial seven tools, a new graphical interface and extended with a new automated query/submission functionality. Thanks to this populating workflow, the SPROUTS database will grow continuously as it is queried.
The SPROUTS database and workflow are freely available at http://sprouts.rpbs.univ-paris-diderot.fr/ as part of the Ressource Parisienne en Bioinformatique Structurale (RPBS) a collaborative portal devoted to structural bioinformatics.

E02: Babak Sokouti, Farshad Rezvan and Siavoush Dastmalchi. Improving the "per residue" prediction accuracy of helical transmembrane segments of GPCRs using GPCRTOP v.1.0 web server

Abstract: The ability to effectively predict the structures of G Protein-Coupled Receptors (GPCRs) is at the heart of helical transmembrane (TM) segments identification. Currently, GPCRHMM is the only GPCR-specific hidden Markov model among different non-specific TM prediction methods. It is designed to determine the TM regions and TM lengths of target GPCRs based on their raw sequences. Besides, there are various classification and identification methods for GPCRs such as PredGPCR, 7TMGPCR, and GPCRsClass. When

considering the prediction capability of a machine learning method, one should carefully give attention to the accuracy measurements (i.e., per protein, per segment, per residue) and the correct implementation of overlapping criteria. However, GPCRHMM suffers from at least two weaknesses. First, it is not able to predict the GPCRs when high values (i.e., 18 and 20) of overlapping criteria are used, as almost 85% of GPCRs have equal or greater than 20 residues at their helical transmembrane segments. Second, its performance at "per residue" accuracy measurement is even lower than other general TM prediction methods.
We have developed a novel HMM based method for the GPCR topology prediction validated on the Uniprot/SWISS-PROT annotated database. The method is based on a 112-state HMM model without considering any physicochemical properties and multiple alignment methods consisting of extra-/intracellular loops (infinite size in length once 15 residues of tails are in place), extra-/intracellular tails (15 residues in length) and helical transmembrane segments (17-25 residues in length). The proposed HMM model benefits from forward algorithm for training process and Viterbi decoding algorithm for predicting the single best path known as the predicted sequence. Therefore, the method considers the transition and emission matrix coefficients for prediction purposes which reveal the grammatical distribution of amino acids along a GPCR sequence. Considering the model of the proposed method, it is able to improve the prediction of GPCR topologies by +3.5% compared to that of the best available method, namely the GPCRHMM. Moreover, it is also capable of recognizing almost 47% of GPCRs when at least 18 residues from each of all seven helical TM segments were predicted correctly, while this accuracy is 41% for GPCRHMM, the next best performing method. We implemented the new method both as a standalone program and as a webserver (GPCRTOP v.1.0; http://biotechnology.tbzmed.ac.ir/?pageid=15). The GPCRTOP v.1.0 is available for users to predict their GPCR sequence of interest by entering its sequence through the textbox and attain their results which include the sequence length, type of protein (i.e., GPCR or Non-GPCR), number of TM segments, and the topology of the GPCR shown as topology state to each sequence position.

E03: Ruben Acuna, Zoe Lacroix, Jacques Chomilier and Nikolaos Papandreou. SMIR: a method to predict the residues involved in the core of a protein

Abstract: Motivation: Protein folding is the critical spontaneous phase when the protein gains its structural con-formation hence its functional shape. Should any error in the process affect its folding the protein structure may fail to fold properly and perform its function. In some cases such misfolded proteins can cause dis-eases. Mutations are typical causes of protein misfolding but some residues are more likely to impact the folding process when mutated than others. This paper presents a new method SMIR that identifies the residues involved in the core of proteins thus more sensitive to mutations.
Results: A Monte Carlo algorithm is used to simulate the early steps of protein folding and the mean number of spatial, non covalently bound neighbors is calculated after 10^6 steps. Residues surrounded by many others may play a role in the compactness of the protein and thus are called Most Interacting Residues (MIR). The original MIR method was updated and extended with a new smoothing method using hydro-phobic-based residue neighborhood analysis. The resulting SMIR method is implemented and available as a server that supports the submission and the analysis of protein structures with MIR2.0 and SMIR. The server offers a dynamic interface with the display of results in a 2D graph.
Availability: The resulting method called SMIR is free and open to all users as a functionality of the Structural Prediction for pRotein fOlding UTility Sys-tem (SPROUTS) with no login requirement at http://sprouts.rpbs.univ-paris-diderot.fr/mir.html. The new server also offers a user-friendly interface and unlimited access to results stored in a database. The SPROUTS database and workflow are freely available at http://sprouts.rpbs.univ-paris-diderot.fr/ as part

of the Ressource Parisienne en Bioinformatique Structurale (RPBS) a collaborative portal devoted to structural bioinformatics.

E04: Surabhi Maheshwari and Michal Brylinski. Prediction of protein-protein interaction sites from weakly homologous template structures using meta-threading and machine learning

Abstract: The identification of protein-protein interactions (PPIs) is vital for understanding protein function, elucidating interaction mechanisms, and for practical applications in drug discovery. With the exponentially growing protein sequence data, fully automated computational methods for predicting interactions between proteins are becoming an essential component of systems-level function inference. The analysis of a representative set of protein complex structures demonstrates that binding site locations as well as the interfacial geometry are highly conserved across evolutionarily related proteins. Since the space of PPIs is highly covered, sensitive protein threading techniques can be used to identify suitable templates for the accurate prediction of interfacial residues for a query protein. Towards this goal, we developed eFindSitePPI, an algorithm that uses the 3D structure of a target protein and evolutionarily weakly related templates to predict binding residues. Using crystal structures, high- and moderate-quality protein models, the average sensitivity (specificity) of eFindSitePPI in interfacial residue prediction is 0.46 (0.92), 0.43 (0.92) and 0.40 (0.91), respectively. This demonstrates that eFindSitePPI is fairly insensitive to structural distortions in modeled target structures. eFindSitePPI also detects specific molecular interactions at the interface; for instance, it correctly predicts approximately one-half of hydrogen bonds and aromatic interactions, as well as one-third of salt bridges and hydrophobic contacts. Comparative benchmarks against various datasets show that eFindSitePPI outperforms other methods for protein binding residue prediction. It also features a carefully tuned confidence estimation system, which is particularly useful in large-scale applications using raw genomic data. eFindSitePPI is freely available to the academic community at www.brylinski.org/efindsiteppi.

E05: Mauricio Macossay Castillo, Simone Kosol, Peter Tompa and Rita Pancsa. Protein structural aspects of multifunctional gene regions

Abstract: Synonymous constraint elements (SCEs) are protein-coding genomic regions with very low synonymous mutation rates believed to carry at least one additional function besides protein coding. Thousands of such potentially multifunctional elements were recently discovered by analyzing the levels and patterns of evolutionary conservation in human coding exons. These elements provide a good opportunity to improve our understanding of how the redundant nature of the genetic code is exploited in the cell. Our premise is that the protein segments encoded by such elements might better comply with the increased functional demands if they are structurally less constrained (i.e. intrinsically disordered). To test this idea, we investigated the protein segments encoded by SCEs with computational tools to describe the underlying structural properties. In addition to SCEs, we examined the level of disorder, secondary structure, and sequence complexity of protein regions overlapping with experimentally validated splice regulatory sites. We show that multifunctional gene regions translate into protein segments that are significantly enriched in structural disorder and compositional bias (low sequence complexity), while they are depleted in secondary structure and domain annotations compared to reference segments of similar lengths. This tendency suggests that relaxed protein structural constraints provide an advantage when accommodating multiple overlapping functions in coding regions.

E06: Pierrick Craveur, Agnel Praveen Joseph, Pierre Poulain, Joseph Rebehmed, Sylvain Léonard, Floriane Noël, Yassine Ghouzam, Romain Deniau, Amine Ghozlane, Jérémy Esque, Guilhem Faure, Aurélie Bornot, Ramachandra Moorthy Bhaskara, Lakshmiparum S. Swapna, Swapnil Mahajan, Garima Agarwal, Vincent Jallu, Jiří Černý, Bohdan Schneider, Catherine Etchebest, Jean-Christophe Gelly, Narayanaswamy Srinivasan and Alexandre G. de Brevern. A short journey inside the protein structures at the light of a structural alphabet

Abstract: Protein 3D structures are directly implicated in the majority of the essential biological functions. To have access to the protein structures can help to understand the precise mechanisms of protein functions. Description of local protein structures have hence focused on the elaboration of complete sets of small prototypes or "structural alphabets" (SAs), that help to approximate every part of the protein backbone. Designing a structural alphabet requires identification of a set of average recurrent local protein structures that (efficiently) approximates every part of known structures. As each residue is associated to one of these prototypes, the whole 3D protein structure can be translated into a series of prototypes (letters) in 1D, as the sequence of prototypes. We had developed a novel structural alphabet with two specific goals (de Brevern et al., 2000): (i) to obtain a good local structure approximation and (ii) to predict local structures from sequence. Named Protein Blocks (PBs), it is the most widely used SA (Joseph et al., 2010).
Two main axes were done using Protein Blocks: (i) methodological developments, (ii) applications to biological systems. Concerning the developments, one of the most interesting ones concerns the protein structure superimposition. Indeed, the 3D structure of a protein can be represented as a sequence of PBs and the alignment of PB sequences effectively reflects the comparison of 3D structures. Both pairwise and structure comparison were assessed better than other available approaches (Joseph et al., 2011, 2012a; Gelly et al., 2011). For instance, an average gain of 84.7% in alignment quality was obtained with respect to the alignments in HOMSTRAD dataset. PBs were used to defined a reduced amino acid alphabet (Etchebest et al, 2007) used since in numerous works. They also were used in various fine analyses of local protein conformation such as cis/trans conformation (Joseph et al, 2012b; Craveur et al, 2013a) or B-bulge conformations (Craveur et al, 2013b). Recently we analyzed protein – protein complexes (Swapna et al, 2012), and protein – DNA interfaces (Schneider et al, 2014).
Protein structures are not static macromolecules, but highly flexible macromolecules. Molecular Dynamics simulation is a good tool to try to handle it. In this field, classical secondary structures assignment is classically used to follow local conformation modification. We have decided to use PBs to analyze them also. We have applied PBs analysis to very different systems: a transmembrane protein named Duffy Antigen / Receptor for Chemokines (de Brevern et al, 2005), camel VHHs (Smolarek al, 2010), proteins with high PolyProline II helix content (Mansiaux et al., 2011, Chevrier et al, 2013), and a particular integrin complex (Jallu et al, 2012). We have underline that some important specificity can only be seen with the use of PBs and not classical secondary structures.

E07: Ludis Morales, Janneth González, George Barreto and David Diaz. Structural and functional predictions of the hypothetical protein PA2481 in Pseudomonas Aeruginosa PAO1

Abstract: Pseudomonas aeruginosa is a bacterium resistant to a large number of antibiotics and disinfectants. Knowing the complete genome sequence including hypothetical proteins, together with encoding processes, provide a lot of information to the discovery and exploitation of new targets for antibiotics. In this study the prediction of the three di-mensional structure of the hypothetical protein, PA2481, from Pseudomonas aeruginosa PAO1 is assessed, and a functional approximation is performed through bioinformatic tools

and software to determine whether it is involved in the antibiotic resistance of this microorganism.

E08: Stanislav Engel and Yosef Kuttner. Misfolded SOD1 noxious "gain-of-interaction" - studying the molecular mechanism of amyotrophic lateral sclerosis (ALS) pathogenesis

Abstract: Amyotrophic lateral sclerosis (ALS) – is a neurodegenerative disease characterized by gradual degeneration and death of the motor neurons. The syndrome belongs to a broad class of so-called "aggregation" diseases (proteopathy), in which a mutation or environmental conditions induces in certain proteins an abnormal conformation (misfolding) characterized by a noxious "gain-of-function". The misfolded protein eventually aggregates but the nature of the toxic intermediate(s) in the aggregation pathway and its mechanism of pathogenesis remain unknown. A prominent ALS pathogenic protein is a ubiquitous enzyme superoxide dismutase 1 (SOD1) which is responsible for a significant fraction of familial and probably sporadic cases. A possible mechanism of misfolded SOD1 pathogenesis is the "gain-of-interaction", in which SOD1 forms aberrant interactions with a variety of cellular proteins, hence interfering with their normal function. The ability to form complexes with structurally diverse proteins is a characteristic of proteins whose surface contains a highly adaptable energetic "hot-spots". The "gain of interactions" of misfolded SOD1 may indicate that some elements of the SOD1's surface acquire certain requisites of the "hot-spots".
In our recent work we hypothesized a role of backbone compliance in protein-protein interactions (PPI) and in the mechanism of binding of small-molecule compounds to protein surfaces. We developed a computational approach of exploring the dynamic properties of protein surfaces using a steered molecular dynamics simulation (SMD). We demonstrated, in a number of model proteins, that a distinct pattern in which static residues form defined cluster(s), the so called "stability patches", surrounded by areas of moderate to high mobility is characteristic of functionally important surface regions involved in PPI and in binding of small-molecule compounds. In the present work we extend the application of the SMD analysis to the field of protein misfolding with the goal of acquiring new insights into the mechanism of noxious "gain of interaction". We demonstrated that upon misfolding, certain areas of the SOD1 surface acquire characteristic properties of the energetic "hot spots", providing a potential explanation for the "gain of interaction" of misfolded SOD1. Identifying SOD1 surface(s) involved in aberrant PPI may facilitate targeted design of small-molecule inhibitors interfering with the formation of pathogenic protein-protein complexes. The fundamental principles underlying the mechanism of transformation of native proteins into noxious ones may be similar for various "aggregation" diseases; therefore, its understanding may pave a way for new strategies to treating these currently intractable diseases.

E09: Michelle Mukonyora. The in silico prediction of foot-and-mouth disease virus (FMDV) epitopes on the South African Territories (SAT)1, SAT2 and SAT3 serotypes

Abstract: Foot-and-mouth disease (FMD) is a highly contagious and economically important disease that affects even-toed hoofed mammals. The FMD virus (FMDV) is the causative agent of FMD, of which there are seven clinically indistinguishable serotypes. Three serotypes, namely, South African Territories (SAT)1, SAT2 and SAT3 are endemic to southern Africa and are the most antigenically diverse among the FMDV serotypes. A negative consequence of this antigenic variation is that infection or vaccination with one virus may not provide immune protection from other strains or it may only confer partial protection. The identification of B-cell epitopes is the key to rationally designing effective high-crossover vaccines that recognize the immunologically distinct serotypes present within the population. Computational epitope prediction methods that exploit the inherent physicochemical

properties of epitopes in their algorithms have been proposed as a cost and time-effective alternative to the classical experimental methods. The aim of this project is to employ in silico epitope prediction programmes to predict B-cell epitopes on the capsids of the SAT serotypes. Sequence data for eighteen immunologically distinct strains from across southern Africa was collated. Since, only one SAT virus has had its structure elucidated by X-ray crystallography (PDB ID: 2WZR), homology models of the eighteen virus capsids were built computationally using Modeller v9.12. They were then subjected to energy minimizations and molecular dynamics using the AMBER force field. The quality of the models was evaluated and validated stereochemically and energetically using the PROMOTIF and ANOLEA servers respectively. The homology models were subsequently used as input to three different epitope prediction servers, namely Discotope2.0, Epitopia and Ellipro. A preliminary set of epitopes has been predicted. In future work, the epitopes predicted in this study will be experimentally validated using mutagenesis studies.

E10: Joseph Rebehmed, Patrick Revy, Guilhem Faure, Jean-Pierre de Villartay and Isabelle Callebaut. The SRI domain family: a common scaffold for RNA polymerase II CTD binding

Abstract: The carboxyl-terminal repeat domain (CTD) of the largest subunit of the RNA polymerase II (RNAPII) serves as docking platform for a wide range of nuclear factors at different stages of the transcription cycle (Egloff and Murphy, 2008). A wide variety of folds are involved in CTD recognition (Meinhart et al. 2005) including a small domain discovered in the histone methylytansferase Set2, called SRI after Set2 Rpb1 Interacting, which alters RNAPII elongation (Kizer et al. 2005). It was recently showed that SRI domain is also present in the C-terminal domain of the RECQ5 helicase, which is critical for maintaining genome integrity (Li et al. 2011). The purpose of this work is to search and study the conservation of the SRI domain during evolution in different protein families and highlight the structural and/or functional features of key residues.
We combined here original tools we previously developed for detecting hidden relationships between remote sequences: (1) SEG-HCA (Faure et Callebaut, 2013b) delineates foldable domains (i.e domains which may form stable 3D structures) from the only knowledge of a single amino acid sequence; (2) TREMOLO-HCA (Faure et Callebaut, 2013a) adds to the results of sequences similarities searches information on domain architecture of the aligned sequences extracted from the Conserved Domain Database (Marchler-Bauer et al. 2013) as well as on the conservation of hydrophobic core residues. I-TASSER threading program (Roy et al. 2010) was used to predict the 3D structure models.
We show that SRI domains are found outside the Set2 and RECQ5 proteins, in which they are also involved in RNAPII CTD or RNAPII CTD-like recognition. Interestingly, SRI domains are always located in the C-terminal extremity of all the proteins of the SRI domain family. The whole family showed a large sequence divergence, especially within the loop between the first 2 helices that is much longer in SET2 proteins than in other members of the SRI family. Despite this divergence that made this domain undetectable by the present SRI CDD profile, it is worth noting that important amino acids for RNAPII CTD binding are highly conserved in the whole family as well as the core hydrophobic residues responsible of maintaining the left-handed three helix bundle fold.
These results allow getting insights into the diversity of this family of domains and into its critical structural and functional features.
This work is supported by a grant from INCa (DIREP).

E11: Yassine Ghouzam, Guillaume Postic, Alexandre G. de Brevern and Jean-Christophe Gelly. Improving remote protein homology detection using a structural alphabet

Abstract: Template-based modeling (TBM) is the most used strategy for protein three-dimensional protein structure prediction. The principle of TBM is to build a structure of a target protein from an experimental structure of related protein (template) using only target sequence. Thus TBM requires to identify and correctly align the best template structure from a database of experimentally resolved structures to the target sequence. Typically sequence alignment methods are used for such task. Nevertheless template selection and target/template alignment remain difficult when proteins are distantly related. Several methods have been specifically developed to increase ability to detect reliable protein templates. Addition of structural information, such as secondary structures, has been shown to improve detection of distantly related proteins. Indeed, proteins might have structural similarities even when no evolutionary relationship of their sequences can be detected while structure is three to ten times more conserved than sequence (llergård et al, Proteins 2009).
We present ORION, a new approach that relies on a better description of local protein structure to boost distantly protein detection. We make use of Protein Blocks (de Brevern et al, Proteins 2000) to improve characterization of local structures. Protein Blocks (PB) encodes a structural alphabet defined by 16 local conformations that describe accurately protein structures. Contrary to secondary structures usually composed of only 3 states (helix, strand and coil), the PB can fairly approximate the local conformation by catching all transitions in protein structures. Starting from a query sequence, a sequence profile is derived using PSI-BLAST and then subsequently used for PB prediction using LOCUSTRA software (Zimmermann and Hansmann, J. Chem Inf Model 2008). Sequence profile and predicted structure of the target are then combined to search in a fold library of profiles built from the HOMSTRAD database (Mizuguchi et al, Proteins Science 1998). We assessed our method on this library and compare it to HHsearch (Söding, Bioinformatics 1998), one of the best method for remote protein homology detection. Our results show that ORION systematically outperforms HHsearch on different benchmarks and detects around 10% more templates at fold and superfamily level. We also show that our method is able to detect more homologs than HHsearch when used on target sequences from the last three CASP competitions. Our method works particularly well for distantly related proteins due to the addition of accurate local structural information.

E12: Olga Kalinina, Jennifer Herrmann and Rolf Mueller. Novel mechanism of S. aureus RNA polymerase inhibition by disciformycins from P. fallax, discovered through structural modeling

Abstract: Multiresistant pathogenic bacteria represent a growing challenge for anti-microbial treatment of nosocomial infections. Thus, there is an urgent need to develop novel antibiotics against them. Amongst other natural sources for these antibiotics, myxobacteria have proven to be highly valuable since they produce a variety of secondary metabolites with antimicrobial activity and unique modes of action [1,2]. Disciformycin, a novel macrolide-glycoside antibiotic from Pyxidicoccus fallax AndGT8, inhibits the growth of multiresistant S. aureus strains in the low micromolar range [3]. In vitro resistance development of S. aureus N315 (MRSA) against disciformycin and whole-genome sequencing revealed several single-nucleotide polymorphisms (SNPs) in rpoB and rpoC that encode for the β and β' subunits of RNA polymerase, respectively; and no cross resistance between disciformycin and several standard antibiotics, including rifampicin, has been observed.
In this study, we have developed a computational pipeline to model the structure of the S. aureus RNA polymerase complex using MODELLER [4] and assess the binding energy of different disciformycins via docking using FlexX [5]. We show that all the SNPs that confer resistance towards disciformycin cluster on the protein-protein interaction interface between the β and β' subunits, thus defining a novel binding site for this small molecule. Docking

shows that binding in the wild-type complex is significantly tighter than in the resistant mutants in accordance with the experimental data. The novel identified binding site differs from the binding sites of other antibiotics, again in accordance with the lack of cross reactivity. Based on these data, we propose a novel mechanism of RNA polymerase inhibition, according to which the inhibitor binds on the protein-protein interaction interface thus precluding the formation of the functional RNA polymerase complex. We note that the proposed methodology is extendable to interpret a broad range of resistant mutations in pathogenic bacteria and other species.

1. Müller R, Wink J, Int. J. Med. Microbiol. 2014, 304, 3.
2. Weissman KJ, Müller R, Nat. Prod. Rep. 2010, 37, 1276.
3. Surup F et al., Angew. Chem., submitted
4. Eswar N, et al., Curr Protoc Bioinformatics. 2006 Supplement 15, 5.6.1.
5. Kramer B, et al. Proteins 1999 37, 228.

E13: Nina M. Fischer, Marcelo D. Polêto, Anders Gärdenäs, Daniel S. D. Larsson and David van der Spoel. Analyzing RNA structures and molecular dynamics simulations

Abstract: The total number of RNA structures available in the Protein Data Bank (PDB) [1] has more than doubled the past ten years. Although, only about 200 non-redundant high-quality (< 2 Å resolution) RNA structures have been solved so far [2], the increase of RNA structures provides more insight at the atomic level and a broader knowledge base for modeling RNA structures.

Due to the growth of three-dimensional structural information and important methodological advances, molecular dynamics (MD) simulations of RNA systems have become more and more significant. MD simulations facilitate to explore important conformational changes, structural flexibility, and solvation processes in detail as well as to refine modeled RNA structures. Using MD simulations for studying RNA structures encounters specific challenges related to their structural features. Some of which are not accurately solved yet, e.g., treating interactions between RNA and metal ions.

The current trend when simulating RNA systems veers toward longer time scales and larger systems. State-of-the-art techniques enable MD simulations of, e.g., whole virus systems including their genome. These simulations create new challenges among others detailed structural analyses. For investigating RNA structures, we developed an RNA analysis tool included within the GROMACS simulation package [3]. With this tool we can analyze PDB structures and also GROMACS MD simulation trajectories of RNA molecules. It extracts information about specific base pairs and structural characteristics similar to DSSR, a new component of the 3DNA suite of software programs [4]. Additionally, we can analyze MD simulation trajectories and thereby, detect changes in RNA base pairing during dynamics as well as stable regions. We present the output not only in text format, but also comprehensibly visualized to gain more clarity. We use our tool to yield statistics of structural features of RNA PDB structures and also to analyze in detail long simulations of virus RNA on the structural level.

[1] Berman et al.,The Protein Data Bank, Nucl. Acids Res., 28(1):235-242, 2000.
[2] Leontis et al., Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. In: Leontis and Westhof, RNA 3D Structure Analysis and Prediction, Vol. 27, 281-298, Springer Berlin Heidelberg, 2012
[3] Pronk et al., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics, 29(7), 845-854, 2013.
[4] Lu et al., 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat. Protoc. 3(7), 1213-1227, 2008.

E14: Marcos Tadeu Geraldo, Agnes Alessandra Sekijima Takeda, Antônio Sérgio Kimus Braz and Ney Lemke. The basis of nuclear import by importin-alpha based on molecular dynamics simulations and normal modes analysis

Abstract: The nuclear import system is responsible for the exchange of protein molecules between the cytoplasmic and nuclear compartments. The most studied system for proteins is the classical nuclear import pathway, which is mediated by importin-alpha and importin-beta with the recognition of a nuclear localization sequence (NLS) motif in the cargo proteins. The aim of our work is to investigate the interaction of NLS and importin-alpha complex using two different computational approaches: molecular dynamics (MD) and normal modes analysis (NMA). We used the NLSs from Ku70 and nucleoplasmin proteins complexed with importin-alpha as our models of study. The MD simulations were clustered based on the RMSD (root-mean-square deviation) values of alpha-carbon atoms. It was generated four and three clusters for Ku70 and nucleoplasmin, respectively, and for each cluster, a reference structure was selected and used for generating the normal modes. Posteriorly, cross-correlation analysis (CCA), principal component analysis (PCA) and the frequency of salt bridges formation were used to analyze the generated data. The CCA indicated similar conformational movement for both simulation techniques, showing a similar contribution of residues in the analyzed complexes (positive correlation of at least 0.75). The PCA showed that PC1 described more than 70% of the conformations generated along the MD and it was possible to correlate it with the first low-frequency normal modes. Differences in the techniques were only found in the salt bridges formation, however, we were able to observe a higher frequency of specific salt bridges, with emphasis on lysine and arginine residues of NLSs interacting with importin-alpha, in agreement with the experimental data described in previous research. We conclude that this computational approach is satisfactory for the understanding of the interaction of protein complexes and we believe it could also be useful in other contexts.

E15: Edda Kloppmann, Burkhard Rost, Jonas Reeb and Michael Bernhofer. Target selection and data analysis for the New York Consortium of Membrane Protein Structure (NYCOMPS)

Abstract: NYCOMPS is a structural genomics center that started in 2005 as part of the Protein Structure Initiative (PSI) with the aim to increase the structural coverage of transmembrane proteins. By the time the PSI will come to an end in 2015, NYCOMPS will have cloned and expressed nearly 15,000 polytopic alpha-helical transmembrane proteins. A unique achievement and experimental dataset. Here, we present the bioinformatics aspects of target selection. In the PSI:Biology phase that began in 2010, NYCOMPS' main focus has been on human homologs. Selected targets are cloned and expressed in the NYCOMPS pipeline and subsequently purified and crystallized, resulting in several unique and novel transmembrane protein structures determined by X-ray crystallography. However, attrition rates are high at every experimental stage. Analyzing the unique data generated in the NYCOMPS pipeline, we attempted to determine features relating to experimental success. We succeeded in developing a prediction method that, using NYCOMPS expression data, prioritizes targets with respect to crystallization success. Furthermore, we present a novel transmembrane helix prediction method, TMSEG, that has been developed as part of the effort to improve target selection.

E16: Eugenia Polverini and Valeria Gherardi. Inside the mechanism of SMN-SmD1 protein complex formation: effects of the Spinal Muscular Atrophy - causing E134K mutation. A molecular dynamics simulation study.

Abstract: Spinal muscular atrophy (SMA) is a motor neuron disease that leads to muscle atrophy due to motor neurons degeneration. SMA is a major genetic cause of early childhood mortality and results from mutations in the Survival of Motor Neuron (SMN) gene1. The SMN protein plays a crucial role in the assembly of spliceosomal small nuclear ribonucleoprotein complexes via binding to the spliceosomal Sm core proteins, in particular to their arginine-glycine (RG) rich C-terminal tails. SMN contains a central Tudor domain, directly involved in the SMN–Sm protein interaction by the recognition of symmetrically dimethylated arginine (DMR) residues in the RG repeats. In particular, an aromatic cage on Tudor domain seems to mediate this binding (1–3).
Six of the pathogenic mutations causing SMA occur in the SMN Tudor domain. The only one that prevents the binding to the Sm proteins without a perturbation of the domain fold is E134K, that is the cause of the more severe type I SMA (3).
To gain more understanding about the mechanism by which SMN interacts with the Sm proteins, and which are the structural effects on binding of its deleterious mutation E134K, we investigated the behavior of the native and mutated structure of the SMN Tudor domain in the presence of the C-terminal tail of SmD1, by means of molecular dynamics simulations.
The interaction of the SmD1 tail with the Tudor domain is electrostatic driven by the acidic residues near the entrance of the aromatic cage. A central DMR of the tail enters into the cage rapidly and stably, forming a network of cationic-pi interactions, both in stacking and T-shaped. The complex is stabilized also by the salt-bridges formed by the other DMRs and arginine residues wrapped around the acidic surface of the domain.
The E134K mutation destabilizes the cage, not only with the disruption of the strong 134-136-127 H-bonds network, but also with the formation of new electrostatic and cationic-pi interactions. The cage collapses and expands, preventing a stable binding of the DMR. This is impeded also by the detachment of the C-terminal region of the tail from the Tudor domain, caused by the E134K charge inversion.
The results are in agreement with what experimentally observed (1–3) and clarify the key role of E134 in the interaction of the SmD1 tail to the Tudor domain. The loss of a strong Tudor-SmD1 interaction, if by one side causes the loss of a functional splicing machinery, by the other side causes the exposition of the detached Sm tails, that could stimulate the recognition by anti-Sm autoantibodies, as is reported for other deseases as lupus erithematosus (4), giving rise to the innovative hypothesis of SMA as an autoimmune desease.
1. P. Selenko et al., Nat. Struct. Biol. 8, 27–31 (2001).
2. R. Sprangers et al., J. Mol. Biol. 327, 507–520 (2003).
3. K. Tripsianes et al., Nat. Struct. Mol. Biol. 18, 1414–20 (2011).
4. H. Brahms et al., J. Biol. Chem. 275, 17122–17129 (2000).

E17: Nicholas Furnham, Natalie Dawson, Christine Orengo and Janet Thornton. Coupling Similarities In Enzyme Reactions With The Evolution Of Their Function Across 375 Domain Superfamilies

Abstract: Many individual and most large-scale analyses of enzyme functions have approached the problem though the study of conservation and divergent relationships between sequences and, where known, their respective atomic structures. In this study we combine this classical approach with newly developed methods to explore the relationships in reaction chemistry that the enzymes perform.
A protocol based on the pipeline described in the construction of FunTree was used to generate phylogenetic trees for 375 structurally defined enzyme superfamilies as defined by the CATH classification. Structurally informed multiple sequence alignments, from which the phylogenetic trees where constructed, were built using a novel agglomerative clustering method. Each tree was annotated with functional information from the UniProtKB resource.

Using the IUBMB reactions describing the function, each reaction was compared to each other using the EC-Blast algorithm. Briefly, this uses atom-atom mapping to derive knowledge of bond changes and reaction patterns for all known biochemical reactions using a variation of the Dugundji-Ugi matrix model. Comparisons were made using three types of normalized similarity scores: bond order, compares the changes in the different number and type of bonds that are being broken and formed; reaction centre compares the local chemical environment around the center of the reaction, and small molecule sub-structures using a common sub-graph detection algorithm that identifies similar fragments. Functional annotations, combined with the phylogenetic tree, were used to infer the ancestral function at each node in the tree using the discrete ancestral character estimation algorithm with an equal rates model as implemented in the APE package in the R statistical suite. This permits the functional changes from parent node to a child node to be traced through the tree and compared using the EC-Blast algorithm. In addition for a subset of 101 superfamilies where catalytic site and structural information is known, the changes in catalytic residue composition and location were automatically analysed and compared to the changes in reaction chemistry. The primary results are displayed via the FunTree web pages (http://cpmb.lshtm.ac.uk/FunTree). Analysis of all superfamilies allows us to present a matrix showing the frequency of exchanges between cataloged functions as well as the frequency of gains/losses of bond types in evolution. Specific examples demonstrating the different extremes of functional evolution can be identified. Large scale structural analysis allows us to highlight the changes in catalytic machinery that have resulted in a shift in function, as well as the surprising number of examples where different catalytic machinery between results in a conservation of function. These observations can be used in addressing a range of biological problems including function prediction, drug development and enzyme design.

E18: Olga S. Voitenko and Olga V. Kalinina. Tight clusters of extremely conserved and non-conserved positions co-localize with protein-protein interaction interfaces of HIV-1 intra-virus and virus-host interactions

Abstract: Proteins are involved in diverse interactions with other molecules within a living cell or in a virus particle. These interactions include interactions with other proteins, ligands, DNA/RNA. Protein-Protein Interactions (PPI) interfaces mediate the specific bindings between these molecules. The principles of protein interactions in general are still poorly understood, and to discover the rules that govern the specificity of protein binding is a fundamental and challenging problem.

Conserved protein residues play important role for maintaining the structure and function of the protein, and have been shown previously to cluster in PPI interfaces and protein active sites [1]. In this work, we investigate the role of clusters of conserved positions in PPIs, in which HIV-1 proteins participate. On the one hand, HIV-1 represents a very appealing study subject, since a lot of sequences and structural data for it are already available, and their amount is constantly growing. On the other hand, most recent studies focus on higher organisms and do not consider inter-organism (e.g. virus-host) interactions.

We use the sequence data from the Los Alamos HIV-1 database [2], and calculate the conservation scores [3] from a multiple sequence alignment created using MAFFT [4] or MUSCLE [5]. We notice that PPIs are located in areas with extreme conservation compared to mean conservation of protein surface (much higher or much lower).

We use an unsupervised learning technique for detecting clusters of residues that have extreme conservation by analogy with the image segmentation approaches: input data are the surfaces of protein 3D structures, and the residues conservation scores are the analogs for pixel intensities. We combined a modified tight clustering approach [6] with spectral and hierarchical clustering, using a distance measure that forces highly conserved, or non-

conserved residues group with each other. We use the Floyd–Warshall algorithm to calculate distances as the shortest path connecting two residues in the residue interaction network on protein surface.
The computed clusters co-localize with observed PPIs in the experimentally resolved complexes of HIV-1 proteins with each other or with the host factors. The obtained clusters are stable over a broad range of parameters. We plan to extend this analysis on other viruses and compare results for different viral families.
[1] Guharoy M, Chakrabarti P (2010):, BMC Bioinformatics, 11:286
[2] http://www.hiv.lanl.gov
[3] Valdar W (2002): Proteins, 48:227-241.
[4] http://mafft.cbrc.jp/alignment/software/algorithms/algorithms
[5] http://www.ebi.ac.uk/Tools/msa/muscle/
[6] Tseng G, Wong W (2005): Biometrics, 61:10-16

E19: Gulin Ozcan, Zeynep Kutlu Kabas, Onur Sercinoglu and Pemra Ozbek. Binding Behavior of HLA-B Alleles Related to Ankylosing Spondylitis (AS) Disease: A Comparative Study by Computational Methods

Abstract: Human Leukocyte Antigens (HLA) are highly specialized proteins forming stable complexes with peptides. Polymorphisms occurring within HLA molecules are associated with various diseases. Ankylosing Spondylitis (AS), which is an autoimmune disease affecting the axial skeleton, is associated with B*27 allele of Human Leukocyte Antigen (HLA). While HLA-B*27:05 is associated with AS, B*27:09 is not associated with AS. These two alleles differ at amino acid position 116 which is located at the peptide binding groove. HLA-B*27:05 has an ASP while B*27:09 has a HIS at this position. Both alleles interact with T cells differently although sharing the same peptide repertoire. Hence, it is of considerable interest both from structural and immunological points of view to understand the differences in binding behavior. In this study, molecular docking and molecular dynamics simulations are carried on aiming to comprehend the differences in the binding behavior of both alleles. Initially, a ‘re-docking’ experiment is applied on both alleles by Autodock 4 for the validation of the molecular docking protocol. This validated protocol is then performed on a library of modeled peptides formed upon single point mutations aiming to address the effect of 20 naturally occurring amino acids at the binding core peptide positions. The free binding energies (FBEs) obtained from computational docking experiments are compared within the peptide library and between the alleles. The amino acid preferences of each position are studied enlightening the role of each on binding. Based on the amino acid preferences of each position, 9 individual peptides are constructed by changing a single position while keeping the rest of the peptide fixed. Molecular dynamics simulations are performed on the docked structures to see the effect of the mutations on the root mean square deviation (RMSD) and cross-correlations.

E20: Grzegorz Chojnowski, Tomasz Waleń, Pawel Piatkowski, Wojciech Potrzebowski and Janusz M. Bujnicki. BrickworX builds models of low resolution nucleic acid crystal structures from recurrent motifs

Abstract: Non-coding RNAs (ncRNAs) were found to be involved in many cellular processes ranging from the gene transcription regulation to the catalysis of chemical reactions. Many ncRNAs, including cis-regulatory elements, form compact, functional, three-dimensional structures that largely determines their function. The method-of-choice for studies of macromolecules structure is x-ray crystallography. However, the RNA crystallography, unlike the protein crystallography, still lacks the methodology facilitating the crystal structure

determination process. In particular software that automatically builds a model into an experimental electron-density map is markedly less developed for nucleic acids than for proteins. In addition the available polynucleotide model building tools require both phosphate and base positions to accurately determine backbone conformers of a single-stranded fragment of a model. This is a limitation since detection of bases is in general much more difficult than phosphates. In particular at low resolution.
BrickworX is a computer program that builds crystal structure models using recurrent RNA motifs extracted from RNABricks database (http://iimcb.genesilico.pl/rnabricks) or B-DNA double helices. In a first step phosphates are detected in a user-provided electron-density map using a Support Vector Machines classifier trained on a large set of crystal structures from PDB. Next, the motifs are positioned in the unit cell based uniquely on the phosphate group pattern. Since fitted motifs comprise at least six bases, the procedure has a relatively large tolerance to missing or wrongly assigned phosphate groups. Nevertheless a fraction of false positives must be filtered out based on low values of the real-space correlation coefficient. To enhance specificity of this step we exploit the base-pairing isostericity properties of nucleic acids. Finally, the most promising trial positions and sequence variants of the motifs are merged to build an output model.
Extensive benchmarks based on a set of over 100 calculated electron-density maps showed that our approach yields models of better quality and covering larger fractions of target structures than other available tools (Nautilus, Arp/Warp). In particular at low-resolutions, below 3.0 Å.

E21: Gwénaëlle André-Leroux, Stéphanie Petrella and Claudine Mayer. Peptidomimetics Based Inhibitor Design for Tuberculosis

Abstract: Among all known bacterial targets, M. tuberculosis DNA gyrase is a validated target for anti-tubercular drug discovery [1]. It has been shown that inhibitors of this enzyme are also active against non-replicating mycobacteria, which is important for the eradication of persistent organisms. A novel inhibitor of M. tuberculosis DNA gyrase would be effective against fluoroquinolone-resistant and multi-drug resistant TB.
DNA gyrase belongs to the bacterial type II topoisomerase family. It is unique in catalyzing the negative supercoiling of DNA. It is assembled from two subunits, each consisting of two structural domains, the N-terminal breakage-reunion (BRD) and the carboxy-terminal domains (CTD) for the A subunit, the ATPase followed by the TOPRIM domain for the B subunit [2]. The double-stranded DNA cleavage site is located in the catalytic core complex composed of the BRD and the TOPRIM domains. The CTD domain is responsible for the negative supercoiling activity. The energy required for catalysis is provided by ATP hydrolysis in the ATPase domain.
We have determined the crystal structures of the four M. tuberculosis DNA gyrase domains [3,4,5]. In the crystal structure of the BRD domain, the N-terminal helix occupies the DNA cleavage site of a symmetry-related molecule [3]. This interaction provides useful starting points for a peptidomimetics approach, a drug designing strategy in which an inhibitor is designed by mimicking the framework of a short peptide. The design of peptide-based inhibitors that target the catalytic site of DNA gyrase shall pave the way for developing unique and original anti-tuberculous drugs.
Coordinates of the helix – BRD complex served as in silico starting template to explore both the binding capacity of the gyrase domain and of the helix. In order to design optimized ligands that could enhance binding and account for the inhibition of the DNA breakage activity, helix residues were extensively mutated. Single, double and multiple cross-mutations were designed for the ligand helix. Wild-type complex as well as variants were optimized with a 5000 iterations protocol using CHARMM force field and the smart minimizer

algorithm implemented in DiscoveryStudio©. After minimization, the interaction energy between the optimized ligand and the protein was evaluated using the Interaction Energy implemented in DiscoveryStudio©. The variants were ranked and compared to wild-type system with a scoring function implemented from total potential energy, interaction energy as well as number of salt bridges and H-bonds interactions. According to these ranking, best mutants were proposed to rational mutation. Experimental tests are currently in progress.

E22: Gabriele Marchler, Farideh Chitsaz, Myra Derbyshire, Noreen Gonzales, Marc Gwadz, Fu Lu, James Song, Narmada Thanki, Josie Wang, Roxanne Yamashita, Chanjuan Zheng, Steve Bryant and Aron Marchler-Bauer. Annotation of Structural Motifs in the Conserved Domain Database

Abstract: The Conserved Domain Database (CDD) is a curated collection of multiple sequence alignments that represent ancient conserved protein domains. CDD includes protein domain models curated as part of the CDD project, models obtained from NCBI's Protein Clusters resource and COGs, as well as models from Pfam, SMART, and TIGRFAMs. Recently we introduced a novel class of models labeled structural motifs. These are defined as compositionally-biased and/or short repetitive regions in proteins, which cannot be modeled as functional globular domains conserved in molecular evolution. Structural motifs include transmembrane regions, coiled coils, and short repeats with variable copy numbers. These models let us annotate such regions more efficiently and accurately. In many cases, only a few position-specific score matrices (PSSMs) suffice to annotate more than 90% of known instances of a specific structural motif. Here, we present the development of structure motif models for several repeat types including leucine-rich repeats (LRR), Armadillo (ARM), HEAT, TPR repeats, and zinc-fingers, among others.
This research was supported by the Intramural Research Program of the National Library of Medicine, NIH.

E23: Zeynep Kutlu Kabas, Gulin Ozcan, Onur Sercinoglu and Pemra Ozbek. Computational Study on the Effect of pH on the Binding Behaviour of HLA-B Alleles

Abstract: HLA subtypes B*2705 and B*2709 differ only in residue 116 (Asp versus His) within their peptide-binding caves, however they are differentially associated with inflammatory rheumatic diseases such as Ankylosing Spondylitis (AS). B*2705 occurs in AS patients, whereas B*2709 is only rarely encountered. Both alleles share the same peptide repertoire, but they interact with T cells differently. Pair wise comparison of the functional, biochemical and biophysical features of these very closely related subtypes may thus illuminate the mechanisms underlying the different disease association. Acidic pH markedly raises association rate constants but dissociation rates are almost unchanged in the pH range 5.0 – 7.0. This pH effect can be described by the protonation/deprotonation states of Histidine. In this study, molecular docking and molecular dynamics simulations are carried on aiming to comprehend the differences in the binding behaviour of both alleles. Initially, peptides extracted from macromolecule were reservoired negatively and positively in order to resemble different pH conditions. On both alleles, computational re-docking is applied by Autodock 4 for the validation of the molecular docking protocol. As a result of docking experiments the free binding energies (FBEs) are obtained and compared within the peptide library and between the alleles for varying pH conditions. Additionally, molecular dynamics simulations are performed on the docked structures to see the effect of the mutations on the root mean square deviation (RMSD) and cross-correlations.

E24: Karolis Uziela, Nanjiang Shu, Björn Wallner and Arne Elofsson. How to select the best protein model using ProQ2?

Abstract: ProQ2 is a model quality assessment program (MQAP) that has been very successful in CASP experiments. The program extracts a number of different input features from the structures of protein models and uses them to train a Support Vector Machine (SVM) that predicts the quality score of the residues in these models. The quality score for the whole model (global score) can be easily derived by summing the scores for each residue (local scores) and dividing the sum by target length.
The global quality scores are important for selecting the best protein model from many possible ones for a given target. ProQ2 is reasonably good at differentiating between very good and very bad models, however, it does not perform well when the task is to pick the best model among many similar good-quality candidates. The goal of this project is to improve ProQ2 performance for this task.
We have tried a number of different strategies to improve ProQ2 performance in picking the best protein model. Some of the things we tried were introducing new training features, changing the scoring function, changing machine learning method and introducing more training samples. Here we present the results of updated ProQ2 program and compare it against the old version.

E25: Mirco Michel, Sikander Hayat, Marcin J. Skwark, Chris Sander, Debora S. Marks and Arne Elofsson. PconsFold: Improved contact predictions improve protein models

Abstract: Application of maximum entropy statistical methods have shown that the quality of protein contact prediction from evolutionary information can be improved significantly if direct and indirect information is separated, provided a sufficient number of related sequences is
available. It was shown that the contact predictions alone contain sufficient information to predict the structure of many proteins. However, since the first reports there have been improvements in the contact prediction methods. Further there also exist slower, but potentially more accurate methods to model the proteins. Here, we ask how much the final models are improved if (i) improved contact predictions are used and (ii) if Rosetta is used to model the proteins.
Using the same 12 proteins as in the original publication of EVfold the TM-score of the top ranked models are improved by on average 33%.
We find that on average the quality is improved with about 15% when the improved contacts from plmDCA are used and another 16% when
PconsC is used. Using Rosetta instead of CNS does not significantly improve global model accuracy. However, the chemistry of models generated with Rosetta is improved.
PconsFold is a fully automated pipeline for ab-initio protein structure prediction based on evolutionary information. PconsFold is based on PconsC contact prediction and uses the Rosetta folding protocol. Due to its modularity, the contact prediction tool can be easily exchanged. The source code of PconsFold is available on GitHub at https://www.github.com/ElofssonLab/pcons-fold under the MIT license. PconsC is available from http://c.pcons.net.

E26: Margot Paulino, Diego Carvalho and Andrés Abin-Carriquiry. New Putative Flavonoid Protein Targets

Abstract: Flavonoids are ubiquitous plant derived compounds broadly recognized by their antioxidant and cytoprotective effects. Although their mechanism of action remains partially explained, growing evidences shows that flavonoids are able to interact with a variety of

protein targets[1]. Recent advances in in silico virtual screening and protein structure prediction allow the identification of new putative protein targets for novel ligands, a strategy known as target fishing[2]. In the present study we propose, taking the flavonoid model quercetin and structurally related compounds, an in silico target fishing approach over proteins with known and unknown crystal structure. The main procedure involves reverse virtual screening based on ligand and protein structure (ChemMapper, IdTarget), ab initio/homology modelling of human homologous proteins (Rosetta, I-TASSER)[3]–[6] as well as analysis and prediction of functionality of new structures and common domains (MOE, ASSIST)[7], [8]. We present a comprehensive analysis of this target space in correlation with ligand structure. The results contribute to flavonoid mechanism of action study and bring new insights into rational drug design.

References

[1] K. Bisht, K.-H. Wagner, and A. C. Bulmer, "Curcumin, resveratrol and flavonoids as anti-inflammatory, cyto- and DNA-protective dietary compounds.," Toxicology, vol. 278, no. 1, pp. 88–100, Nov. 2010.
[2] R. N. Gacche and H. D. Shegokar, "Evaluation of Selected Flavonoids as Antiangiogenic , Anticancer , and Radical Scavenging Agents : An Experimental and In Silico Analysis," pp. 651–663, 2011.
[3] J. Gong, C. Cai, X. Liu, X. Ku, H. Jiang, D. Gao, and H. Li, "ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method," Bioinformatics, vol. 29, no. 14, pp. 1827–1829, Jul. 2013.
[4] J. Wang, P. Chu, C.-M. Chen, and J. Lin, "idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach.," Nucleic Acids Res., vol. 40, no. Web Server issue, pp. W393–9, Jul. 2012.
[5] B. Raveh, N. London, L. Zimmerman, and O. Schueler-furman, "Rosetta FlexPepDock ab-initio : Simultaneous Folding , Docking and Refinement of Peptides onto Their Receptors," vol. 6, no. 4, 2011.
[6] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER : a unified platform for automated protein structure and function prediction," vol. 5, no. 4, pp. 725–738, 2010.
[7] H. 2R7 Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, "Molecular Operating Environment (MOE), 2011.10," 2011.
[8] S. Caprari, D. Toti, L. Viet Hung, M. Di Stefano, and F. Polticelli, "ASSIST: a fast versatile local structural comparison tool.," Bioinformatics, vol. 30, no. 7, pp. 1022–4, Apr. 2014.

E27: Adva Yeheskel, Rony Seger and Malka Cohen-Armon. Protein Complexes Assembly and Function Revealed by Structural Motion Prediction

Abstract: PARP1 (polyADP-ribose polymerase-1) is a DNA binding protein which transfers ADP-ribose from nicotineamide dinucleotide to glutamic and aspartic residues. It contains 7 domains, 4 of them bind DNA and one is a catalytic domain. We explore the relationship between function and dynamics for PARP1 using structural motion prediction (normal mode analysis). A network of inter-domain contacts links nicked DNA binding domains to the catalytic domain in PARP1. Recently, the PARP1 was shown to be activated after binding to phosphorylated ERK2 in the absence of nicked DNA (Cohen-Armon et al.Molecular Cell 2007). The ability of PARP1 to bind nicked DNA is associated with DNA repair, whereas binding phosphorylated ERK2 mediates a variety of physiological functions associated with regulation of gene expression. This includes epigenetic mechanisms and mechanisms promoting proliferations and differentiation. In this work we predict the binding site of PARP1 to phosphorylated ERK2 and demonstrate the allosteric effects implicated in the

catalytic activity of PARP1 using normal mode analysis. We show conformational changes in PARP1, rendering the protein accessible to target proteins binding, which may allow catalytic activity. These conformational changes are maintained either with nicked DNA or with ERK2 bound to PARP1. We suggest that the observed motion underlies PARP1 function, a hypothesis that should be further validated.

E28: Mateusz Banach, Elodie Duprat, Mathilde Carpentier, Barbara Kalinowska, Irena Roterman and Jacques Chomilier. Identification of the folding nucleus of globular protein: application to immunoglobulin-like and flavodoxin fold domains

Abstract: Folding nucleus of globular proteins is a group of interacting amino acids whose reciprocal interactions (mainly hydrophobic) allow the formation and stability of the 3D structure. It can be predicted by simulation of the early steps of the folding process with a Monte Carlo coarse grain model in a discrete space. We previously defined MIRs (Most Interacting Residues), as a set of residues presenting a large number of non-covalent neighbour interactions during such simulation. In order to define the minimal number of residues giving rise to a given fold instead of another one, MIRs can be considered as good candidates, although their proportion is rather high, typically 15-20% of the sequence length. Having in mind experiments with two sequences of very high levels of sequence identity (up to 90% and above) but different folds, one can admit that our method slightly over estimate the residues coding for the fold. In order to better predict it, we combined the MIR method, which takes sequence as single input, with the "fuzzy oil drop" (FOD) model, which requires a 3D structure. FOD assumes that a globular protein follows an idealised 3D Gaussian distribution of hydrophobicity density, with the maximum in the centre and minima at the surface of the "drop". If the actual local density of hydrophobicity around a given amino acid is as high as the ideal one, then this amino acid is assigned to the core of the globular protein, and it is assumed to follow the FOD model.
In this study, we compared and combined MIR and FOD methods to define the minimal nucleus of two populated domain folds: immunoglobulin-like and flavodoxins. Each 3D structure dataset (56 Ig-like and 37 flavodoxins, respectively) is composed of distantly related domain sequences, therefore with certain diversity in the functions, but with a common fold. The combination of these two approaches defines some positions both predicted as a MIR and assigned as highly hydrophobic according to the FOD model. It is shown here that these methods, using different methodologies based on different folding states of the protein domains (early step of the folding, simulated from amino acid sequences by MIR method, and final state as resolved in the 3D structure and compared to an ideal Gaussian function by the FOD method), are in agreement, as indicated by the significant intersection of their key predicted residues, for both Ig-like and flavodoxin folds. In most cases, sources of any discrepancy may be explained on the basis of protein structural specificity. The performance of joint prediction improves the fitness of the putative folding nucleus determination by significantly decreasing the number of positions capable of forming high levels of hydrophobic interactions, to the level of 5% in each fold. It provides a reasonable agreement with positions experimentally demonstrated as belonging to the common folding nucleus of Ig-like protein domains.

E29: Marc Lensink and Shoshana Wodak. Score_set: A CAPRI Benchmark for Scoring Protein Complexes

Abstract: Motivation: The CAPRI protein-protein docking experiment has proven to be a catalyst for the development of docking algorithms. An essential step in docking is the scoring of predicted binding modes in order to identify stable complexes. Since 2005, the CAPRI

experiment has been providing enriched data sets, including both correct and incorrect docking solutions, to enable developers to test new scoring functions independently from docking calculations. Here we present an expanded benchmark data set for testing scoring functions, comprising the ensemble of predicted complexes made available for scoring calculations in the CAPRI experiment.
Results: The expanded scoring benchmark contains predicted complexes for 15 published CAPRI targets. These targets were subjected to 23 CAPRI assessments, due to existence of multiple binding sites for some targets. The benchmark contains more than 19000 protein complexes. About 10% of the complexes represent docking predictions of acceptable quality or better, the remainder representing incorrect solutions (decoys). The benchmark set contains models predicted by 47 different predictor groups, including web servers, that use different docking and scoring procedures, and is arguably as diverse as one may expect, representing the state of
the art in protein docking.
Availability: The data set is available at the following URL: http://cb.iri.univ-lille1.fr/Users/lensink/Score_set

E30: Yannick Spill and Michael Nilges. Variance and Information Content in SAS profiles

Abstract: Small-Angle X-Ray or Neutron Scattering (SAS) are experiments which provide low-resolution structural information on biological macromolecules. They require a monodisperse solution of the molecule of interest, and can be performed at biologi- cally relevant temperatures. Among other quantities, SAS gives access to the radius of gyration, whether the object is elongated or globular, whether it is flexible or not. The Debye formula (Debye P, Ann. Phys., 1915) describes the link between interatomic distances and SAS profile, the product of a SAS experiment. Using this formula, a recent paper (Moore P. B., Biophys. J., 2014) showed that thermal motion strongly affects the SAS profile of a protein in solution, especially at wide scattering angles. Previous work studied the variance and covariance of a SAS profile (Spill, PhD Thesis, 2013). Here, we show that these results convey a much deeper understanding of the structure of the data provided by a SAS experiment. We show that the information contained within a SAS profile is not spread out uniformly. Instead, information is most abundant at intermediate angles. Study of the correlation structure of SAS experiments also shows that different resolution levels are independent in the SAS profile. This independence calls for specific treatments when scoring the agreement between two profiles. Overall, this work is a major step towards understanding what can and cannot be obtained from a SAS profile.

E31: Emilio Potenza, Tomas Di Domenico, Ian Walsh and Silvio Tosatto. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins.

Abstract: Intrinsically disordered regions are key for the function of numerous proteins. Due to the difficulties in experimentally characterizing disorder, many computational predictors have been developed with various disorder flavors. Here we provide a new version of MobiDB, a centralized source for data on different flavours of disorder in protein structures covering all UniProt sequences (over 45 million). The database features three levels of annotation: manually curated, indirect and predicted. The new version also features a consensus annotation for long disordered regions. MobiDB aims at giving the best possible picture of the "disorder landscape" of a given protein of interest. Manually curated data is currently extracted from the DisProt database. Indirect data is inferred from missing residues in PDB experiments that are considered an indication of intrinsic disorder. Deposited files of NMR experiments for protein structure resolution often contain multiple models. By

calculating the differences between the positions of each model's residues, one can measure the degree in which this positions change and interpreted as a measure of how flexible or disordered a protein is. Since MobiDB currently covers the full set of UniProt sequences, the included predictors need to be extremely fast. The ten predictors currently included (ESpritz in its three flavours, IUPred in its two flavours, DisEMBL in two of its flavours, GlobPlot, VSL2b and JRONN) enable MobiDB to provide disorder annotations for every protein, even when no curated or indirect data is available. In order to complement the disorder annotations, MobiDB features additional annotations from external sources: Annotations from the UniProt database include post-translational modifications, and linear motifs. Pfam annotations are displayed in graphical form and are link-enabled, allowing the user to visit the corresponding Pfam page for further information. Secondary structure is extracted from the PDB whenever available, and displayed in graphical form.
MobiDB was designed with a multi-tier architecture, using separate modules for data management, data processing and presentation functions. To simplify development and maintenance, all tiers handle the common JSON (JavaScript Object Notation) format, thereby eliminating the need for data conversion. The MongoDB database engine is used for data storage and Node.js as middleware between data and presentation. MobiDB exposes its resources through RESTful web services, by using the Restify library for Node.js. The Angular.js framework and Bootstrap library were selected to provide the overall look-and-feel. MobiDB offers users both graphical web interface access and RESTful web services from URL:http://mobidb.bio.unipd.it/. Crosslinks to MobiDB can also be found in each UniProt page inside the "3D structure databases" section.

E32: Olga Zanegina, Anna Karyagina, Andrei Alexeevski and Sergei Spirin. Structural classification of protein-DNA complexes and their families based on interacting elements.

Abstract: Systematization of protein-DNA complexes is necessary for understanding mechanisms of protein-DNA interaction. As the number of new structures rises every year, it is important to create an approach for automatic classification of new protein-DNA complexes.
In current work we considered structures of SCOP protein domains in contact with double-stranded DNA that were extracted from PDB entries. Suggested classification is based on the interacting structural elements of protein and DNA. For each complex we determined all contacts between the DNA molecule and the protein domain. As contacts we consider both hydrogen bonds and hydrophobic interactions. Each contact was assigned to a secondary structure element of protein (α-helix, β-strand, or non-structured) and to one of the following elements of DNA: the major groove, the minor groove, or the backbone. Thus any contact belongs to one of types (e.g. “helix with tha major groove” or “strand with backbone” etc.). As a result we characterized each domain–DNA complex by a set of contact types observed in it. Interaction type of a SCOP family was defined as a list of those contact types that were represented in all domains of the family. The variety of all observed interactions of structural elements in a family could help to precise contacts in a particular complex and to level crystallization artifacts caused by minor protein motions.
We analyzed protein-DNA contacts in 794 structures from 118 families of SCOP protein domains in contact with DNA. All variety of observed sets of contact types was divided onto 37 interacting groups. In most cases several interacting groups were found within the same SCOP family. At the moment there are 29 SCOP families containing three or more different proteins whose structures are solved in contact with double-stranded DNA. These families are divided onto 10 interaction types.
Suggested classification can be extended to newly arrived PDB structures automatically.
Acknowledgements: this work was supported by RFBR grant 13-07-00969.

E33: Maciej Antczak, Tomasz Zok, Martin Riedel, David Nebel, Piotr Lukasiak, Marta Szachniuk, Thomas Villmann and Jacek Blazewicz. Accurate approach for nucleotide conformation prediction of RNAs

Abstract: Knowledge of RNA 3D structure is critical for understanding the numerous functions that RNAs play in living cells. RNA 3D structure prediction is often solved by ab initio or homology modeling approaches, although remains a difficult challenge. Quality of both approaches can be improved when an accurate method of nucleotide conformation prediction onto fixed backbone coordinates will be applied. In contrast to problem of protein side-chain conformation prediction, there are no available tools to solve a corresponding problem for RNAs. To fill this gap, we propose a computational method that allows to reconstruct high-quality, full-atomic structure of RNA using backbone-dependent library of discrete nucleotide conformations (called rotamers) based on backbone coordinates obtained from homological or artificial structure. Moreover, this approach can also be used to refine incomplete structures. It supports the following prediction modes: bases only and ribose rings and bases. Proposed solution integrates two components. First, a backbone-dependent library of RNA rotamers which is used to identify best fitting nucleotide conformations for every input residue. Second, an optimization algorithm that allows to find a low-energy conformation in a reasonable time. Rotamer libraries were constructed by machine learning approaches (e.g. kmedoids, neural gas), which classify nucleotide conformations upon backbone torsion angles, ribose puckering and glycosidic bond of every nucleotide observed in experimentally determined high-resolution structures (<2.4Å) stored in RNA FRABASE. Conformation energy is computed taking into consideration rotamer probabilities and non-local steric atom-atom interactions. The optimization algorithm integrates dead-end elimination procedure and graph theory approach tailored to our needs to identify molecule conformation with the global minimum energy. The resultant structure energy is minimized in the Cartesian atom coordinate space using NAMD2 program with CHARMM force field. The method assures short processing time and high quality of predicted RNA 3D structures. We conducted the evaluation test for 40 RNAs of different structural complexity with strand length ranged from 30 to 161 nt. The high accuracy of prediction is reflected by both Root Mean Square Deviation (RMSD) and Interaction Network Fidelity (INF) measures, computed by RNAlyzer, between the predicted and crystal structures. The average values computed for the entire input set of 40 RNAs was for RMSD: 0.826Å and 0.935Å and for INF: 0.971 and 0.927, for bases only or ribose ring and bases prediction mode respectively. An analysis applying MolProbity tool shows high quality of all predicted structures. Around 90% of ribose puckering and glycosidic bond angles are predicted correctly within 30° of the X-ray counterparts, as measured by MCQ4Structures. We conclude, that proposed approach proves to be useful in the process of structure reconstruction.

E34: Sayoni Das, David Lee, Natalie Dawson and Christine Orengo. FunFHMMer : Exploiting CATH-Gene3D functional families to predict functions and functional sites of uncharacterised sequences

Abstract: Due to the rapid increase in international genome-sequencing initiatives and structural genomics projects, a large amount of protein sequence and structural data are accumulating. Since experimental characterisation of such huge amounts of data is not feasible, computational approaches that can predict protein functions and functional sites are essential.
We propose a domain based method for predicting protein functions that exploits functional subclassification of superfamilies in CATH-Gene3D. CATH-Gene3D superfamilies have

been subclassified into functional families using a hierarchical agglomerative clustering algorithm supervised by a family identification protocol, FunFHMMer, that recognises highly conserved positions and specificity-determining positions in cluster alignments and uses this information to ensure functional coherence. The functional purity of the families has been assessed using a set of manually curated mechanistically diverse enzyme superfamilies and an in-house residue enrichment analysis based on the percentage of conserved residues in a family that coincide with experimentally determined functional residues. The families were further associated with Gene Ontology terms probabilistically, in order to predict functions for uncharacterised sequences.
FunFHMMer outperforms our previous method DFX, which was the top domain-based method in the first Critical Assessment of Function Annotation (CAFA) experiment. Moreover, the function annotations provided by FunFHMMer families are shown to be more precise than annotations provided by Pfam. CATH currently identifies ~110,300 functional families and for the most populated of these (accounting for 72% of CATH-Gene3D sequences), residues implicated in functional sites can be predicted.

E35: Wim Vranken, Daniele Raimondi and Elisa Cilia. Applying dynamics-based interaction potentials in a residue network

Abstract: Accurate information on the dynamics of proteins is difficult to obtain, which hampers investigations of their importance in relation to protein function. We recently developed DynaMine [1,2] (http://dynamine.ibsquare.be), which provides predictions for the fast backbone movements of proteins directly from their amino-acid sequence. These predictions are based on a linear regression model derived from NMR chemical-shift derived backbone dynamics information [3] for proteins in solution.
We have now trained and analysed twenty different per-residue linear regression models to derive how amino acids affect each other' backbone dynamics. The underlying data reflects the most thermodynamically stable state of the protein in solution and covers proteins from folded to disordered. We therefore propose that the strength of the determined amino acid interactions reflects their statistical propensity to stabilize particular conformations (reduced dynamics). Furthermore, the interactions are directional, meaning that the effect of a particular amino acid (e.g. tryptophan) on another (e.g. an alanine following it in the sequence) is not the same as the reverse (e.g. the effect of an alanine on a tryptophan preceding it). We also trained corresponding linear regression models by randomization of the input sequences in order to filter out spurious amino acids effects. We further extended the above analyses to secondary structure populations derived from NMR chemical shifts using the δ2D method [4].
Finally, we explore the use of these statistical dynamics and secondary structure interactions in protein-specific residue networks, similar to previous work based on three-dimensional structures [5]. These network representations are built starting only from the protein primary structure. Since they give a detailed picture on how amino acids influence the dynamics of other residues along the sequence, they complement the information provided by the DynaMine dynamics profile and might be useful in the future to engineer protein sequences with desired dynamical characteristics.
1. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. From protein sequence to dynamics and disorder with DynaMine. Nat Commun 4, 2741 (2013).
2. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. The DynaMine webserver: predicting protein dynamics from sequence. Nuclei Acids Res (2014).
3. Berjanskii, M. V. & Wishart, D. S. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. Nucleic Acids Res 35, W531-537 (2007).
4. Camilloni, C., De Simone, A., Vranken, W. F. & Vendruscolo, M. Determination of

secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 51, 2224–2231 (2012).
5. Konrat, R. The protein meta-structure: a novel concept for chemical and molecular biology. Cell Mol Life Sci 66, 3625–3639 (2009).

E36: Elisa Cilia, Rita Pancsa, Peter Tompa, Tom Lenaerts and Wim F. Vranken. DynaMine: a web-server for predicting protein dynamics from sequence

Abstract: Dynamics are an essential aspect of protein function. Nevertheless, understanding them poses significant challenges, mainly because accurate protein dynamics information remains difficult to obtain.
We have recently released the DynaMine web-server [1], which provides predictions for the fast backbone movements of proteins directly from their amino-acid sequence. DynaMine rapidly produces a dynamics profile describing the statistical potential for the backbone movements at residue-level resolution.
By exploiting the statistical analysis of NMR data of proteins in solution DynaMine gives quantitative insight into the relationship between amino acid sequence and backbone dynamics, as we have shown in [2].
The webserver underlining predictors consist of linear regression models trained on backbone N-H S2 order parameter values for 210880 residues in 1952 proteins; these S2 values are estimated with the Random Coil Index [3] from a carefully assembled dataset of NMR chemical shift data extracted from the BioMagResBank (BMRB) [4]. S2 order parameters represent how restricted the movement of an atomic bond vector is with respect to the molecular reference frame, with physical values varying between 1 for fully restricted (rigid conformation) and 0 for fully random movement (highly dynamic).
DynaMine takes as input a protein sequence and produces a profile of per-residue predicted S2 values (S2pred). Each S2pred is predicted based on the local sequence environment provided by the 25 residues preceding and following the target residue in the sequence. Sibling predictors that consider a shorter sequence context enable the webserver to provide predictions also for short peptides. In addition, we have determined ranges of predictive values where a residue is likely to be rigid, flexible, or has highly context-dependent dynamics. The prediction results and the dynamics profiles annotated according to this analysis are visualized and can be directly downloaded.
Through this webserver, we aim at providing molecular biologists with an efficient and easy to use tool for predicting the dynamical characteristics of any protein of interest, even in the absence of experimental observations. The DynaMine webserver, including instructive examples describing the meaning of the profiles, is available at http://dynamine.ibsquare.be.
1. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. The DynaMine webserver: predicting protein dynamics from sequence. Nuclei Acids Res (2014).
2. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. From protein sequence to dynamics and disorder with DynaMine. Nat Commun 4, 2741 (2013).

E37: Amrita Roy Choudhury, Marjana Novič and Igor Zhukov. The transmembrane regions of Bilitranslocase

Abstract: The goal of our work is to elucidate the transport channel structure and functional mechanism of the transmembrane protein bilitranslocase. The primary function of bilitranslocase is to transport organic anions, including bilirubin, anthocyanins, nicotinic acid and other micronutrients. Bilitranslocase shows a wide range of tissue distribution and is potentially druggable. To analyze the protein structure, we have used a combination of computational and experimental methods.

Bilitranslocase has four transmembrane alpha helical regions. The transmembrane regions TM2 and TM3 are playing key roles in forming the transport channel, in ligand binding and mediation. These two transmembrane regions also overlap with the two ligand binding sites of the protein. The probable assembly of the four transmembrane regions is analyzed with Monte Carlo approach. Predicted interhelical interactions between transmembrane regions TM2:TM3 and TM1:TM4 serve as the primary constraint during the sampling. Analyzing the best-scoring conformations indicate three probable assemblies of transmembrane regions. In the most populated assembly, the two key transmembrane regions, TM2 and TM3, are arranged diagonally opposite to each other. In addition, the structures of these two regions are analyzed, both individually and in mixture, with NMR experiments performed in SDS micelle environment.
The transmembrane regions TM2 and TM3 constitute of amino acids participating in H-bond formation, which play key role in ligand-bilitranslocase interactions, and are flanked by ligand binding motifs. Further, structures of both the transmembrane regions show Pro induced kinks, which render flexibility to the transport channel. These structural features are in line with the metastable nature of the protein. Present work is focused on understanding the dimensions of the transport channel combining results from computational and experimental studies.

E38: Amine Ghozlane, Etienne Ruppé, Julien Tap, Nicolas Pons, Alexandre De Brevern, Joseph Rebehmed, Sean Kennedy and Stanislav Ehrlich. Pairwise Comparative Modeling For Identification Of Class A Beta-lactamases In the Human Intestinal Microbiota

Abstract: Background: Metagenomics have revolutionized the apprehension of complex microbial environment. However, the functional annotation based on sequence homology remains challenging. Indeed, the low identity shared by some predicted proteins with known proteins hampers their assignation to a family. Herein, we propose a new 3-dimensional approach, pairwise comparative modeling (PCM) consisting in 3-D homology modeling of a candidate protein using two templates matching the two closest protein families. PCM was applied to the identification of class A beta-lactamases (BLAA) in the Human intestinal microbiota (MetaHIT study).
Methods: Potential BLAA were searched in a non-redundant protein catalogue (n=3.9M) using a BLAA reference dataset (n=676) and the Hmmer, Blastp and SSearch softwares. Candidates were modeled using Modeller with in parallel (i) BLAA templates and (ii) low-molecular weight protein-binding proteins (PBP, homologs to BLAA) templates. Besdies, a set of 60 BLAA and 42 PBPs were submitted to the same process. Modeling and alignments yielded various scores that were used to build a logistic regression that was subsequently applied to discriminate BLAA from PBP among the candidates.
Results: We identified 212 putative BLAA candidates which displayed at least 25% identity with any BLAA. Only 18 (8.5%) had >95% identity with reference BLAA. Most of the candidates shared identity with BLAA recovered from functional metagenomic studies (38.7%) or with BLAA from anaerobic bacteria (36.8%). Conventional annotation (eggNOG v3) could assign 194 (91.5%) BLAA and 17 (8.0%) PBP. PCM and eggNOG were concordant for the identification of BLAA and PBP in 166 (78.3%) and 16 (7.5%), respectively. However, PCM assigned a PBP function to 28 (13.2%) candidates that would have been assigned a BLAA function by eggNOG. Among those 28, structural alignment showed that 25 (89.3) had no glutamate close to reference position glutamate 166, essential to the BLAA function.
Conclusions: PCM was consistent with eggNOG for BLAA positive identification, but also appeared to be more specific as it invalidated a BLAA assignation given by eggNOG in 13.2% cases. PCM is a promising approach in metagenomic investigations.

E39: Tunca Dogan, Alex Bateman and Maria Martin. UniProt Domain Architecture Alignment: A New Approach for Protein Similarity Search using InterPro Domain Annotation

Abstract: Motivation: Similarity based methods have been widely used in order to infer the properties of genes and gene products containing little or no experimental annotation. The most popular ones are the sequence similarity search methods such as BLAST. New approaches that overcome the limitations of the methods that relying solely upon sequence similarity are rising. One of these novel approaches is the comparison of the organization/architecture of the structural domains in the proteins. The idea is that the shared structural units may indicate shared evolutionary and functional properties associated between these units.
Results: Here we propose a new algorithm for the comparison of domain architectures in order to identify similarities and to propagate functional annotations between the proteins in the UniProt Database. The method "UniProt Domain Architecture Alignment" is unique from previous approaches in three major ways: (i) the use of InterPro Database for the domain annotation, (ii) the incorporation of the domain weights into the dynamic programming step, and (iii) the inclusion of information regarding non-annotated regions in the proteins into the domain architectures. The performance of the method was measured through the identification of orthology using the OMA database (F1 score: 0.62). The results indicated the effectiveness of the approach for similarity detection. We plan to integrate the algorithm into a learning based system for the automatic annotation of uncharacterized proteins in the UniProtKB/TrEMBL database.

E40: Benjamin Bardiaux, Barth van Rossum, Olivera Francetic, Nadia Izadi-Pruneyre, Christiane Ritter, Hartmut Oschkinat and Michael Nilges. Structural modelling of symmetric protein assemblies from distance constraints.

Abstract: Dealing with symmetric oligomeric structures is an important issue in the context of structural biology since it is estimated that about 60% of the proteins in every genome are homo-oligomers. While X-ray crystallography is still the most important tool for protein complexes three-dimensional structure elucidation, some very important biological machines are refractory to crystallization, like membrane proteins or large symmetric assemblies such as bacterial pili or amyloid fibers.
Nuclear magnetic resonance spectroscopy (NMR), particularly in the solid-state, is also been used for collection of high-resolution structural data, in the form of spatial proximities. Other resonance spectroscopy approaches (FRET and EPR) can also provide structural data encoded as distance constraints. More recently, chemical cross-linking coupled to mass spectrometry has emerged as a powerful technique to study the structure and organization of large biomolecular assemblies by detecting pairs of residues close in space.
In this context, support from molecular modelling is crucial for translating distance data into three-dimensional structures. Here, we present a general method, based on strict symmetry relations, for structure calculation of high-order symmetric aggregates from distance constraints. The approach is not limited in terms of symmetry group and thus finds direct applications for elucidation of oligomer, membrane protein or fibril structures. We will illustrate the method through several examples, from small oligomers studied by solution NMR to large helical assemblies and combining different type of structural data.

E41: Yuezhou Zhang, Alexandre Borrel, Leslie Regad, Anne-Claude Camproux, Gustav Boije Af Gennäs, Jari Yli-Kauhaluoma and Henri Xhaard. Phosphate and ribose structural isosteric replacement in the Protein Data Bank

Abstract: In this work, we have studied the structural isosteric replacements of phosphate and ribose found in protein-ligand complexes available in the Protein Data Bank database (June of 2014 release) [1]. We developed a computational protocol that was used to construct 157 datasets. Each of these datasets is composed of several superimposed ligands, including POP, AMP, ADP or ATP to use as references, derived from superimposed molecular complexes. Structural replacements of ribose and phosphate groups were then extracted and studied: we identified a set of 15 common structural isosteres of phosphate and 43 structural isosteres of ribose. In addition to classical isosteres of phosphate, we found unexpected types of replacements that do not conserve charge or polarity, for example phosphate and ribose replaced by aliphatic groups, phenyl, or carbamoyl groups. The structural mechanism involved in structural isosteres appears varied: New interactions may be created, water molecules are important, in some case ion plays a role, and of course large and small conformational changes do occur at the binding sites. This study has implications both in the field of medicinal chemistry [2-4], i.e. it expands our knowledge of structural isosteres, and in the field of chemoinformatics, since our results have implications with respect to the definitions of chemical similarity.

[1] Berman,H.M. et al. (2000) The Protein Data Bank. Nucleic Acids Res., 28, 235–242.
[2] Elliott, T. S., Slowey, A., Ye, Y., & Conway, S. J. (2012). The use of phosphate bioisosteres in medicinal chemistry and chemical biology. MedChemComm.
[3] Papadatos, G., & Brown, N. (2013). In silico applications of bioisosterism in contemporary medicinal chemistry practice. Wiley Interdisciplinary Reviews: Computational Molecular Science, 3(4), 339–354.
[4] Hamada, Y., & Kiso, Y. (2012). The application of bioisosteres in drug design for novel drug discovery: focusing on acid protease inhibitors. Expert Opinion on Drug Discovery.

E42: Bjoern-O. Gohlke, Robert Preissner, Tim Overkamp, Antje Richter, Bernd Gillissen and Peter Daniel. Target landscapes identifies Vatalanib as PARP inhibitor

Abstract: In the present study, we used two-dimensional (2D) and three-dimensional(3D) structural similarities comparisons followed by generation of target similarity landscapes to identify new functions and/or side-effects of known bioactive compounds directed against twelve crucial anti-cancer targets like VEGFR, EGFR, PDGF or PI3. Therefore, we utilized more than 10,000 compounds from our SuperTarget database with known inhibition values for different anti-cancer targets and performed an all-against-all comparisons resulting in 2D and 3D similarity landscapes. Interestingly, there are regions with low 2D but high 3D similarity scores, which we examined in detail. This detailed analysis showed the unexpected structural similarity between inhibitors of vascular endothelial growth factor receptor (VEGFR) like Vatalanib and inhibitors of poly ADP - ribose polymerase (PARP).
By using an in silico docking simulation and an in vitro PARP assay we confirmed that the VEGFR inhibitor Vatalanib exhibits off - target activity as a PARP inhibitor, broadening its mode of action.
Thus, in contrast to the 2D-similarity search, the 3D-similarity landscape comparison identifies new functions and side effects of the known VEGFR inhibitor Vatalanib.

E43: Alexandre Borrel, Leslie Regad, Henri Xhaard, Michel Petitjean and Anne-Claude Camproux. Druggability prediction performances related to different pocket estimations

Abstract: Therapeutical molecules bind to preferred sites of action, which are in the majority located within proteins or at their surface. Therefore, estimation and characterization of pockets is a major issue in drug target discovery. Among the molecules, “drug-like

molecules" are small molecules with particular properties as able to cross the digestive tract. Pocket druggability, the ability of a pocket to bind "drug-like", is essential for drug discovery studies [2] especially for discovering new targets.
Identifying druggable pockets is possible by different statistical models of prediction [2-5] which differ in methods used to estimated pockets, in descriptors used to characterized pockets and in statistical methods used. Moreover, the quality of these approaches is limited by the few data available, and by the pocket estimation [6], pockets changes with the estimations. However of new target discovery, it is important to be able to predict the "druggability" of a pocket in its apo form that means when it is not yet bound to a ligand and deformed by the interaction with one ligand.
We propose a model to predict pocket druggability from holo or apo form. From one protein set [5], we used different approaches to estimate pockets. With ligand, defines pockets as protein atoms less than 4 Å away from the ligand or without ligand, based on two geometric estimations DoGSite [7] and fpocket [6]. Pockets estimated using three approaches, were characterized, using 57 descriptors [9]. Three pocket sets are generated.
From each pocket sets, we built a pocket "druggability", based on a Linear Discriminant Analysis (LDA). The construction of these models consisted in the selection of LDA models with the best accuracy and containing as few descriptors as possible. The model, giving the best performance is conserved. Finally we used a consensus of 4 LDA models, built from pockets estimated by fpocket and which present a good accuracy (close to 80%), better than literature [3-6], on other pocket sets and an apo pocket set [5].
[1] Nisius, B., Sha, F., & Gohlke, H. (2011). Journal of biotechnology, 159(3), 123–134
[2] Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., & Rarey, M. (2012). J. Chem. Inf. Model., 52(2), 360–72.
[3] Desaphy, J., Azdimousa, K., Kellenberger, E., & Rognan, D. (2012). J. Chem. Inf. Model., 52(8), 2287–99
[4] Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S., Brenk, R. (2011). J. Chem. Inf. Model., 51(11), 2829–42
[5] Schmidtke, P., & Barril, X. (2010). Journal of Medicinal Chemistry, 53(15), 5858–67
[6] Pérot S, Sperandio O, Miteva MA, Camproux AC, Villoutreix BO (2010). Drug Discovery Today, 15, 656—667.
[7] Volkamer, A., Kuhn, D., Rippmann, F., & Rarey, M. (2012). Proteins, 81
[8] Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009) BMC bioinformatics, 10, 168
[9] Pérot,S., Regad,L., Reynès,C., Spérandio,O., Miteva,M.A.,Villoutreix,B.O., Camproux,A.C. (2013). PLoS One 8[6], e63730

E44: Jairo Rocha, Ricardo Alberich and Emidio Capriotti. DRFLEX: An RNA Structural Classification Database with RNAFlex

Abstract: Motivation. The RNA database DARTS is already 6 years old and no longer updated.
However, there are an increasing number of unclassified RNA structures in the PDB that need to be compared and studied to better define their biological activity. A number of programs have been developed for alignment but most of them assume that the structures are rigid and penalize the alignments according to the RMSD of the alignments.
Method. We created a new classification tree of RNA structures using the new program RNAFlex. We have implemented the program RNAFlex, (manuscript in preparation), that evaluates the similarity between two RNA structures using a sequence of local transformations, instead of a single rigid transformation for the entire matching. A dynamic programming strategy optimizes the alignment matching single bases and basepairs

evaluating the rigid local transformation of structural neighbors. The scores associated to all possible pairs are used to generate a classification tree that defines the RNA structural space. Results. Since structure and function are highly correlated, we validated our approach on a set of functionally annotated RNA structures against the ARTS, SARA and SETTER programs. We show that RNAFlex results in an overall improvement of the classification performance and significant overlap with previously mentioned method. Finally we built a new RNA classification database that is compared to DARTS.

E45: Pietro Lovato, Alejandro Giorgetti and Manuele Bicego. A multimodal approach to protein remote homology detection

Abstract: Protein remote homology detection (PRHD) represents a crucial and challenging task in bioinformatics, with different promising solutions already appeared in the literature. Even if reaching satisfactory accuracies on several benchmark datasets, there are still complex cases where even these state-of-the-art approaches may perform poorly: in such cases, it seems reasonable to try improving the unsatisfactory results by incorporating and exploiting other types of information. From a general point of view, this may lead to a multimodal approach, i.e. an approach aimed at solving a given task by integrating different sources of information. In the context of PRHD, there is a source of information which is typically disregarded by classical approaches: the available experimentally-solved, possibly few, 3D structures. In this paper we provide some evidence that such information can be successfully integrated in a system of PRHD. Our approach is inspired by the multimodal image and text retrieval context [1], where images are equipped with loosely related narrative text descriptions, and are retrieved using textual queries. This scenario is particularly interesting with respect to our scopes, because of the following similarities: i) the link between the modalities is weak, partially hidden, and in general difficult to infer; ii) the context is asymmetric: one of the two modalities is richer than the other, yet being more difficult or expensive to obtain - therefore fewer examples are typically available. The goal is to develop an approach which works directly on the weaker source of information (the text), being however built taking into account the smaller richer source (the image).
In this paper we propose a novel approach to PRHD, which exploits the afore-described multimodal point of view: in particular, the richer modality is represented by a subset of 3D structures retrieved from PDB. The proposed approach starts by encoding sequences and structures with a count representation, namely a representation obtained by counting the number of occurrence of some basic elements inside an object: sequences are described using counts of Ngrams, structures are described using counts of 3D fragments. Both representations are then modeled using topic models, a class of probabilistic approaches for count data: in particular we investigated two models, the Latent Dirichlet Allocation (LDA) and the Componential Counting Grids (CCG) model - creating in the end a "structure-aware" model for sequences.
Various tests on the SCOP 1.53 benchmark demonstrate that our approach can improve results in those scenarios where the sequence modality fails, even when a very reduced amount of structures are available. A further detailed analysis on a GPCR protein confirms that this multimodal approach can extract information that cannot be obtained from sequence-based techniques.
[1] Jia et al., "Learning Cross-modality Similarity for Multinomial Data", ICCV 2011

E46: Deepti Jaiswal, Radka Svobodova Varekova, David Sehnal, Crina-Maria Ionescu, Stanislav Geidl, Lukas Pravda, Vladimır Horsky, Michaela Wimmerova and Jaroslav Koca. Consistency of sugar structures and their annotation in the PDB

Abstract: Cell-cell recognition is the first stage in many important phenomena such as infection by bacteria and viruses, communication among cells of lower eukaryotes, binding of sperm to egg, etc. [1]. Cell-cell recognition relies on sugar (carbohydrate) specific interactions at the cell surface. Theoretical studies typically involve molecular modeling of sugars and sugar-specific protein receptors. These studies rely on structural information obtained mainly by crystallography and nuclear magnetic resonance, and deposited in the Protein Databank (PDB). Since the main purpose of PDB is to store the structure of proteins and nucleic acids, thus, it is expected that PDB structure files are complete and correctly annotated. Nonetheless, sugars exhibit a structural diversity larger than amino acids or nucleotides, a property which makes them ideal for recognition. At the same time, sugars are characterized by specific and very sensitive structural features such as multiple chiral centers on each ring. Because of these peculiarities, the validation and annotation of sugar structures is not straightforward.
Our first goal was to develop a methodology that can identify whether a sugar structure is complete and correctly annotated. Our second goal was then to check all PDB entries containing sugars, and record whatever problems we encounter in the sugar structures. For this purpose we collected all sugar structures which appear as ligands in PDB entries, and compared them to model structures available in Ligand Expo [2], a curated repository of ligand chemical and structural information. In order to perform the comparison we used several tools for structural comparison currently available (SiteBinder [3], Open Babel [4]), as well as two in-house programs. We report here on our findings regarding the complete and correctly annotated sugar structures in PDB, together with the problematic cases.
References:
[1] Brandley BK, Schnaar RL: Cell-surface carbohydrates in cell recognition and response.1986 Jul, 40(1):97-111.
[2] Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J: Ligand Depot: a data warehouse for ligands bound to macromolecules. Bioinformatics 2004 Sep, 1;20(13):2153-5.
[3] Sehnal D, Vařeková RS, Huber HJ, Geidl S, Ionescu CM, Wimmerová M, Koča J: SiteBinder: an improved approach for comparing multiple protein structural motifs. J Chem Inf Model 2012 Feb, 27;52(2):343-59.
[4] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: Open Babel: An open chemical toolbox. Journal of Cheminformatics 2011, 3 :33.

E47: Nicolas Denis and David Ritchie. Fine-Grained Structure-Function Clustering of Pfam Protein Domain Families: A Case Study Using CYP450

Abstract: It is well known that protein structures are more evolutionarily conserved than their sequences. However, until recently it has been very computationally expensive to compare multiple protein structures using only their three-dimensional (3D) shapes. While current protein structure databases such as SCOP (Murzin et al., 1995) and CATH (Orengo et al., 1997) provide extremely useful resources, these databases have several limitations. For example, it is generally very expensive to search them using structure-based queries, and they require considerable manual effort to keep them up to date (Holm et al., 2008).
In order to help tackle such problems, we have recently developed a new protein structure alignment and comparison tool called “Kpax” (Ritchie et al., 2012; http://kpax.loria.fr). This allows us to perform all-against-all structural alignment and comparison calculations rapidly and reliably. However, despite its speed, Kpax is limited to performing only structure-based comparisons. While this allows rapid clustering of protein domain shapes, we believe it would be very useful to be able to group and classify protein domains using other attributes and functional annotations as well as fine-grained differences in 3D protein shape. As a first step

towards this goal, we are studying the protein domain structures of the cytochrome P450 enzyme family. According to the Pfam classification (Finn et al., 2010), this domain family (Cyp450) currently has some 1023 structures in the Protein Data Bank (PDB), of which 554 are structurally non-redundant after duplicate chains and domains are removed.
We have developed a series of Python scripts to extract automatically the annotations available for these structures from the UNIPROT and InterPro web sites, for example, and to perform a variety of shape-based and shape-function clustering calculations using R scripts. We find that Cyp450 can be clustered into some 22 structural groups, and these groups seem consistent with the 5 main functional classes listed on InterPro. We are currently analysing in detail the functional annotations of these groups, and we are working to extend our approach to be able to apply it automatically to all of the structural domains of all Pfam families.
References
Finn,RD, et al. (2010). The Pfam protein families database. Nucleic Acids Research, 38, D211–D222.
Holm, L, et al., (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics, 24, 2780-2781.
Murzin, AG, et al. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 247, 536-540.
Orengo, CA, et al., (1997). CATH - A Hierarchic Classification of Protein Domain Structures. Structure, 5, 1093-1108.
Ritchie, DW, et al. (2012). Fast Protein Structure Alignment using Gaussian Overlap Scoring of Backbone Peptide Fragment Similarity. Bioinformatics, 28, 3278-3281.

E48: Jad Abbass and Jean-Christophe Nebel. Customised fragment libraries for ab initio protein structure prediction: usage of functional and structural annotations

Abstract: When template-based modelling cannot be applied on a target sequence, fragment-based protein structure prediction is currently the best ab-initio alternative. Those methods rely on constructing putative protein structures by iterative stochastic selection of fragments from a database of known structure elements. Although a specific set of fragment libraries is designed for each prediction software, they are independent from the protein of interest. In this study, it is proposed to produce customised libraries by incorporating functional and structural annotations of the target sequence. Under that scheme, only protein structures sharing the annotations, i.e. Gene Ontology functions, CATH or SCOP classes, with the target are selected to construct fragment repositories. Using the open-source Rosetta package, a state-of-the-art fragment-based protein structure prediction software, this novel approach has been evaluated on targets of the latest CASP experiments. Comparison with standard Rosetta predictions demonstrates the added value of creating bespoke fragment libraries.

E49: Irena Roterman, Mateusz Banach, Barbara Kalinowska and Leszek Konieczny. Similar Structure – Different Stabilization – Analysis Of Immunoglobulin-Like Domains

Abstract: “Fuzzy oil drop” model represents a general framework for describing the structure of hydrophobic core in proteins and provides insight into the influence of the water environment upon proteins structure and stability. Model assumes comparison of two hydrophobicity density distributions: idealized and really observed one in particular protein. The idealized hydrophobicity density distribution is assumed to be represented by 3-D Gauss function where the center of the molecule is characterised by the highest hydrophobicity density which decreases according to bell curve as the distance increases and reaching zero level on the surface of the protein molecule. The 3-D Gauss function encapsulates protein molecule in ellipsoid, the size of which is expressed by σ parameters for each axis

independently (in probability calculus - standard deviation). Knowing the parameters of Gauss function the expected hydrophobicity density can be calculated for each point inside the protein molecule. On the other hand the observed hydrophobicity density is the effect of local pair-wise hydrophobic interactions between residues of particular hydrophobicity parameter. Residues are represented by effective atoms (averaged position of all atoms belonging to side chain of particular residue). The Levitt function was applied to calculate the hydrophobic interaction. These two distributions in the protein body can be compared quantitatively using the divergence entropy (Kullback-Leibler distance entropy), the value of which measures the degree of accordance/discordance between theoretical and observed distributions classifying protein molecule as more or less stable due to the regular/irregular structure of hydrophobic core. The divergence entropy can be applied also to evaluate the participation of each secondary structure fragment in the general structure of hydrophobic core. The immunoglobulin-like domains represent similar super-secondary structure – two-layer sandwich, however their biological activity is different (immunoglobulins, enzymes, structural proteins, DNA-binding). The "fuzzy oil drop" model-based analysis of immunoglobulin-like domains reveals different participation of particular Beta-structural fragments in the general structure of hydrophobic core. It may be interpreted as different stability of these fragments and in consequence different ability for structural changes. Immunoglobulin-like domains are known as undergoing structural changes leading to amyloid forms. Thus different participation in hydrophobic core may explain their different ability to be transformed into amyloids. This hypothesis was verified using tranthyterin as an example. Quantitatively measured participation in core formation helps explain the variable stability of proteins and is shown to be related to their biological properties. The goal of the work is to verify the possibility to treat the "fuzzy oil drop" as the possible criteria for amyloidosis prediction.

E50: Aram Gyulkhandanyan. Binding of cationic porphyrins to heme proteins

Abstract: Motivation: Hemoglobin and cytochrome c are the most investigated heme proteins as from a structural point of view, as well as of their functional features. At the same time and those and other studies to date are still in progress, complementing the new knowledge about these proteins. In photodynamic therapy of tumors (PDT) exogenous porphyrins (photosensitizers) in the blood interacts with proteins and transported to tumors. Investigation of protein-porphyrin interaction and competitive binding of other ligands (particularly of fatty acids) for binding sites on proteins of blood are the important tasks for PDT. In the interaction of an array of porphyrins and several possible carrier proteins these tasks are further divided into many tasks on optimization of binding conditions of each pair of protein-porphyrin. One of the new effective methods to solve such tasks of multidimensional biology is the method of small molecule microarrays (SMM). Another method - absorption spectroscopy with high accuracy allow to determine the binding constants of ligands (porphyrins) with proteins. These two independent experimental methods allow to determine the most perspective porphyrins for use in PDT. Investigations by computer simulation method (molecular docking) substantially complement experimental methods, allow to predict the possibility of protein-ligand interactions and to conduct screening of ligand binding sites. This paper considers new properties of these proteins - binding of exogenous cationic porphyrins with hemoglobin and cytochrome c.
Results: By method of small molecule microarray under identical conditions was investigated binding of 30 porphyrins to serum albumin, Hb and cytochrome c. It has been shown that serum albumin and hemoglobin equally well bind cationic porphyrins and that cytochrome c, as and the carrier proteins, binds well with porphyrins. By methods of absorption spectroscopy it has been shown that the highest affinity to cationic porphyrins showed

cytochrome c (1.5 - 2.0 times higher compared to Hb and BSA, respectively). Collectively of the data, that some porphyrins at lower concentrations are stimulators of apoptosis and that the cytochrome c plays an essential role in the mechanism of apoptosis, may assume that the high affinity of cationic porphyrins to cytochrome c is an important experimental fact and that porphyrins can have very essential significance to launch of apoptosis. By the method of molecular docking shown that the main site for the binding of ligands (porphyrins, fatty acids and their complexes) is an internal cavity of the macromolecule Hb, and that complexes [porphyrin-fatty acid] may displace free porphyrins from the internal cavity of the macromolecule of Hb. Molecular docking method is a good and available tool for screening of ligand binding sites on the protein macromolecule and can add significantly the data of the experimental methods about structural features of proteins.

E51: Bedrat Amina, Amrane Samir, Guédin Aurore and Mergny Jean-Louis. Algorithm to predict G-quadruplex folding through score computing

Abstract: Single-stranded guanine rich DNA can form stable secondary structure in vitro called G-quadruplex (G4) DNA. Structurally, G-quadruplex consists of stacks of square arrangement of four guanines (a tetrad or quartet) in planar Hoogsteen hydrogen bonded form. This structure is stabilized by monovalent cation e.g. K+ and Na+. The corners of the tetrads are linked by three nucleic acid sequences (loops) of varying composition and topology. The high thermodynamic stability of G-quadruplexes under near-physiological conditions suggests that these structures may form in genomic DNA in vivo. Approximately 370000 sequences with putative G-quadruplex-forming motifs are dispersed throughout the human genome. Indeed various oncogenes (c-myc , c-kit) have been shown to have G4 motifs in the promoter region G4 structures are associated with a number of important aspects of genome function, which include transcription, recombination and replication.
Bioinformatic approaches have played an important role by identifying new candidate sequences with the potential to form G-quadruplex. The current Bioinformatic tools are based on limited structural information that could be reformulated. Thus the number of putative G-quadruplexes in the genome is expected to be larger than previously reported.
We propose an arbitrary but objective Algorithm that would predict G4 folding propensity from a linear nucleotidic sequence. The new method focuses on Guanines clusters and GC asymmetry, taking into account the whole genomic region.
After transforming the nucleotides to numbers, the algorithm calculates a score for the sequences. The higher the score is, the higher is the possibility of the sequence to form a G4. The accuracy of our prediction Algorithm has already been proved and compared to previously predicted sequences. From our analysis of an aligned 1684 HIV-1 we detected 10 new conserved G4-forming sequences.
The conserved G-rich region from HIV-1 promoter, known to regulate the transcriptions of the HIV-1 provirus, showed its ability to form G4 structure. We determined the structure of the sequence located on the HIV-1 promoter and we also patented this sequence for its potential to inhibit HIV virus at nanomolare range
Finally this algorithm has found new quadruplex putative sequences, which cannot be identified by previous computational methods.

E52: Luigi D'Ascenzo and Pascal Auffinger. Electrostatic potential dissimilarities between aromatic amino acids and nucleobases lead to anion-π stacking events in nucleic acids

Abstract: In order to better characterize RNA folding and RNA structure as well as RNA interactions with proteins and ligands, it is essential to refine our knowledge of the non-covalent interactions that are at play in these systems. Without doubt hydrogen-bonds are

among the best known non-covalent interactions. But, next to them, a large diversity of non-covalent interactions exists. For instance, in the protein world, cation-π interactions, namely the stacking of cationic species over an aromatic group, became rapidly a necessary letter in the non-covalent interaction alphabet. Yet, despite their obvious aromatic character, no significant cation-π interactions were described so far for nucleic acid systems. Here we report that nucleic acid aromatic systems prefer to interact with anionic rather than cationic species.

Indeed, through the calculation of electrostatic potential surfaces, we were able to show that the protein and nucleic acid aromatic systems do not share the same characteristics. The formers are electron-rich systems able to establish cation-π contacts; the latters are electron-depleted systems that display a propensity to attract anions and establish anion-π interactions. This largely unnoticed dissimilarity between aromatic rings of both main biomolecular groups sheds new light on some of their essential properties. Through an exhaustive search of the PDB for anion-π interactions involving, among others, the DNA and RNA backbone phosphate groups, it is found that anion-π interactions are rare in DNA compared to RNA where they are mostly involved in sharp turns (75%) such as those found in tRNA anticodon loops and, more generally, in RNA tetraloops. Besides, these interactions are also observed between sequence-distant residues in ribosomes and in crystal-lattice contacts. The relative rarity of these anion-π interactions outside of loops indicates that it is generally not prevailing over classical hydrogen bonds in the nucleic acid context but nevertheless vital for their folding, structure and function.

E53: Luigi D'Ascenzo and Pascal Auffinger. Unusual neutral Asp/Glu protein side chains interacting with nucleic acid structures

Abstract: Interactions with nucleic acids involving cationic species (mainly Na+, K+ and Mg2+) are relatively well characterized. Conversely, less is known about anionic species interacting with nucleic acids. In the last decade some investigations aimed at clarifying anion RNA interactions and in particular the ones involving Asp/Glu side chains. Favored nucleobase binding sites emerged; more specifically, guanine Watson-Crick sites were observed to dominate over other possibilities without excluding less obvious interaction patterns.

We report here several surprising instances of close contacts between Asp/Glu side chains and hydrogen bond acceptors in nucleic acid systems. A remarkable example involves a contact with a backbone phosphate group that interacts simultaneously with a protonated form of a sulfate anion (PDB code: 2DRB; pH: 7.5). These counterintuitive contacts involving Asp/Glu residues generally considered as negatively charged and nucleic acid electronegative atoms can only be understood by considering the neutral forms of these amino acids. To confirm that these interactions are not anecdotal, a survey of the PDB was performed (res ≤ 2.5 Å, electron density maps checked in all occurrences). This survey collected a large array of short hydrogen bond contacts between Asp and Glu residues with themselves and with other electronegative atoms, therefore stressing that their neutral form is more frequent than previously thought.

The purpose of this study is to highlight the implications of such unexpected hydrogen bond network contributions for the understanding of protein/nucleic acid interactions. As a main outcome it is proposed that the neutral Asp and Glu amino acids have to be included into an expanded bestiary of interacting residues leading to novel recognition patterns. These motifs will certainly have to be considered in forthcoming nucleic acid/protein prediction studies.

E54: Lukáš Pravda, Radka Svobodová Vařeková, David Sehnal, Crina-Maria Ionescu, Karel Berka, Michal Otyepka and Jaroslav Koča. Anatomy of enzymatic channels and algorithm for its detection

Abstract: Biomolecular channels and pores are indispensable for a huge variety of cell-life processes. They enable passage of substrate/product compounds to/from the active site in case this site is deeply buried within the protein structure. Additionally, pores enable transport of variety of molecules through lipid bilayer. Physicochemical properties of these channels such as polarity, hydropathy, charge or bending and radius greatly influence the specificity and selectivity of enzymatic reaction. Therefore, precise detection and, especially characterization of their properties is of a main interest of many researchers involved in rationalizing their roles in enzymatic reactions. Such knowledge can be directly utilized in drug design, rational design of enzymes and other biotechnological application.
We analyzed a set of 4316 enzymes in six enzymatic classes with the known active site from CSA database with the sequence identity lower than 90%. In this dataset we identified that more than two thirds of enzymes contain at least two channels of minimum length 15Å. We found out that the average composition of channels significantly differs from the average composition of proteins, their surface or core. The overall chemical properties of channels also differ among individual enzymatic classes. This is a first time wide-range study of enzymatic channels revealing their importance for co-determining enzyme substrate preference. For the analysis we utilized a successor of the tool well-received by a scientific community for identification of channels MOLE 2.0. This tool is capable of not only fast and interactive analysis of individual small proteins or large protein-nucleic acid complexes, but also large sets of proteins. In this poster a results of this study will be present so as the MOLE 2.0 tool. MOLE is available as a standalone application and plugins for popular molecular-browsers free of charge at (http://mole.chemi.muni.cz) or as web-service (http://mole.upol.cz).

E55: Evgeniy Aksianov and Andrey Alexeevsky. ProtOn: a tool for automatic annotation of beta-structures

Abstract: We developed web-service ProtOn (http://mouse.belozersky.msu.ru/proton) which includes four tools for analysis of 3D structures of all-beta and alpha/beta classes.
1. SheeP is the detector of beta-sheets in an input structure in PDB format. The algorithm has three steps. (A) Secondary structure detector DSSP is used to detect sheets. (B) Maps of all sheets are created. Sheet map is a table, which cells correspond to amino acids, rows - to strands and columns - to sets of amino acids, connected by bridges (in terms of DSSP program). (C) SheeP inspects spatial arrangement of C-alpha atoms and modifies sheet maps: amino acids can be removed or added to a sheet map, sheets can be joined or separated. SheeP was shown to be more accurate sheet detector than DSSP and STRIDE [1].
2. ArchiP is a detector of protein architectures, i.e. spatial arrangements of units - sheets, parts of sheets and adjacent layers of helixes. An architecture is represented by a graph, which vertexes correspond to units and edges – to contacts between them. Using the graph, ArchiP detects a number of common architectures (sandwiches, barrels, propellers and prisms). ArchiP sensitivity in determining common architectures vary from 68% (for barrels) to 99% (for alpha/beta-sandwiches).
3. MotAn is a detector of structural motifs. It represents the topology of polypeptide chain from N- to C-terminus like "[A1-]-[h1]-[A2+]-[B2-]-[B3-]-[h2]-[B1-]". [h1] indicates first helix, [A1-] - a strand in upper raw of the map of sheet A (edge strand of sheet A), [A2+] - strand in the next raw of the map, different signs means that pair of strands [A2+] and [A1-] is antiparalell.
MotAn detects meanders, interlocks, jellyrolls, Greek keys, beta-alpha-beta-motifs and beta

helixes in the created topology. For example, "[Ai+]-[hj]-[A(i+1)+]" denotes beta-alpha-beta-motifs. MotAn was shown to be more accurate than PTGL [2].
ArchiP and MotAn distinguish core and minor units and strands. Core are large ones; minor are smaller. Minor ones can be skipped in the architecture or topology description.
MotAn and ArchiP sensitivity is due to the use of SheeP, which attempted to detect sheets according human judgment, and skipping minor units and strands in the result.
4. ProTop (Protein Topologies) is a beta-version of the tool to align the topologies, created by MotAn and thus, to find the weak similarities of the proteins. For example, "[A1-]-[h1]-[A2+]-[B2-]-[B3-]-[h2]-[B1-]" can be aligned with "[A1+]-[A2-]-[B2+]-[B3+]-[h2]-[B1+]" after inverting strands' direction and inserting the gap after the first strand of the second topology.
The work was partially supported by RFBR grants 14-04-31709 and 13-07-00969.
1. E. Aksianov, A. Alexeevski. SheeP: a Tool for Description of Beta-Sheets in Protein 3d Structures. Journal of Bioinformatics and Computational Biology, 2012, 10(2)
2. E. Aksianov. Motif Analyzer for protein 3D structures. J Struct Biol. 2014. 186(1): 62-67.

E56: Yasaman Karami, Serge Amselem, Elodie Laine and Alessandra Carbone. Disease-related mutations in proteins: a study of dynamically correlated networks and coevolved residue clusters

Abstract: Proteins ensure their biological function by adapting their shapes and motions in response to environmental conditions. This structural plasticity can be altered by genetic modifications (point mutations) that can induce effects at distant protein sites, thereby provoking diseases. Networks of dynamically correlated residues play a crucial role in propagating such perturbation signals across protein structures. These residues are also expected to display high degrees of conservation and/or coevolution. However the relationship between sequence evolution and structural dynamics has been seldom explored yet.
In this study, we performed a consensus analysis of dynamically correlated and coevolved residue networks in three archetypal proteins: tumor suppressor p53 (highly flexible), protein A (highly stable), and in growth hormone which its mutants are involved in human genetic diseases. Dynamically correlated residues were detected from all-atom molecular dynamics simulations and grouped into: (i) Communication Pathways (CP) stabilized by non-bonded interactions and (ii) Independent Dynamic Segments (IDS) displaying concerted atomic fluctuations (Laine et al., PloS Computational Biology, 2012). IDSs and CPs represent two complementary media for allosteric communication and they cover almost all the protein structure. Coevolved residues were detected using BIS (Dib and Carbone, PloS ONE, 2012) and clustered with CLAG (Dib and Carbone, BMC bioinformatics, 2012).
Our results first reveal a significant overlap in all studied proteins between clusters of coevolved residues and networks of dynamically correlated residues. Most of the coevolved residues are involved in IDSs and/or CPs. Moreover clusters of coevolved residues can be classified according to whether they contain residues participating in CPs, in IDs or in both.
Second, we examined the impact of two disease-related mutations on the allosteric communication profile of growth hormone. The comparison of dynamically correlated residue networks in wild-type and mutated proteins showed a rewiring of CPs linking coevolved residues. In particular, more coevolved residues are connected in the mutant compared to the wild-type.
Consequently we have shown that characterizing the dynamical behavior of proteins provides a means for a physical understanding of coevolution signals encoded in sequences. Understanding the role of disease-related mutations on the link between coevolution and dynamical correlation can help decipher the molecular mechanisms of mutation-induced

allosteric deregulation.
E. Laine et al., “Allosteric Communication across the Native and Mutated KIT Receptor Tyrosine Kinase”, PLoS Computational Biology (2012).
L. Dib, A. Carbone, “CLAG: an unsupervised non hierarchical clustering algorithm handling biological data”, BMC Bioinformatics (2012).
L. Dib, A. Carbone, “Protein fragments: functional and structural roles of their coevolution networks”, PLoS ONE (2012).

E57: Nidhi Tyagi, Edward Farnell, Colin Fitzsimmons, Stephanie Ryan, Rick Maizels, David Dunne, Janet Thornton and Nicholas Furnham. Allergenic Proteins Are Targets For Mammalian IgE Mediated Immune Response Against Metazoan Parasites: Linking allergy to immunity against metazoan parasites

Abstract: Allergenicity can be described as an uncontrolled and hostile Type 1 hypersensitive immune reaction mediated by antibody IgE towards environmental antigens from diverse sources such as foods, plants and various organisms. The mechanism responsible for eliciting the allergenic reaction (unregulated immune response) involves components of the immune system that also regulate the immune response against infection of multicellular parasites such as intestinal worms and skin-attaching mites (regulated immune response). In the absence of parasitic infection, this system turns hostile and becomes hyper responsive towards innocuous environmental proteins possibly due to similar molecular features of the two. We seek to explain this peculiar behavior of an otherwise beneficial and highly evolved immune system to allergenic conditions by investigating if significant sequence and structural similarity exist between and parasitic proteins that render immunity to the host against infections. In the quest of understanding relationships between immunity and allergenicity, the systematic exercise of assessing the epitopic regions of allergenic proteins and ‘epitope-like’ regions of parasitic proteins becomes of primary importance. Building a comprehensive knowledgebase of similar/homologous proteins from allergenic sources and parasites through comparison of their salient features would provide further insight to our understanding of the underlying mechanism of onset and progression of allergenic response. There are 10 protein domain families (Tropomyosin, EF hand, CAP, Profilin, Lipocalin, Cupin, Bet v 1, Trypsin-like serine protease, Expansin and Prolamin) that represent nearly 45% of all documented allergenic proteins and thus are highly populated.
We searched 4196 protein sequences that are known to cause allergy against dataset of proteins encoded in genomes of helminths (platyhelminthes (flat worms) and nematodes (round worms)) and parasitic arthropods (mites). By employing various computational approaches, we could arrive at a comprehensive list of parasitic proteins, which show significant sequence and structural similarity with allergenic/IgE binding proteins corresponding to 10 highly populated allergenic protein domain families.
We also searched sequence and structure motifs of known IgE-binding protein fragments (epitope) from allergens in parasitic proteins, we identified epitopic-like regions in 197 parasitic proteins for families that represent allergenic proteins.
We have also performed experimental analysis to establish IgE-binding activity for the parasitic proteins that we predicted by performing computational analysis. We present for the first time, a comprehensive catalog of allergenic proteins mimicking immunogens, which will impact the discovery and design of molecules used in immunotherapy in allergenic conditions. The findings from this study will further enrich our understanding of allergenicity in the light of immunity.

E58: Géraldine Caumes, Hiba Abi Hussein, Jean-Baptiste Chéron, Anne-Claude Camproux and Leslie Regad. Effects of the pocket estimation algorithms in one pocket-ligand complex classification

Abstract: Understanding the mechanism behind drug-side effects is of major interest. One of the most important scenarios seems to be the interaction between the drug and additional targets. The protein-ligand interaction occurs in a preferential protein region, named pocket. Thus, to be able to assess potential side effects of a given drug by predicting which targets this drug can interact with, is an important aspect for drug discovery and development.
In a previous study (Perot et al. PLos One 2013), we investigated the combined pocket and ligand spaces by developing a classification of pocket-ligand pairs extracted from 483 protein-ligand complexes. Pockets were estimated using an energy-based approach (Jain Comput Aided Mol Des, 2007). Each pocket-ligand pair was then characterized using 24 descriptors allowing a description of their geometrical and physico-chemical properties. Using this pocket-ligand pair description, we performed a classification of pocket-ligand pairs, which reveals 5 main clusters with particular pocket and/or ligand properties. An analysis of these clusters shows correspondences between pocket and ligand properties. For instance small pockets tend to be more polar and bind small and polar ligands.
The methods used to estimate pockets from the tri-dimensional structures are classified into 4 categories (Pérot et al., DDT 2010): proximity (Halgren JCIM 2009), energetic-based (Jain Comput Aided Mol Des 2007), geometric-based (Schmidtke et al. NAR 2010) and template-based methods (Chéron et al. in submission). All these methods provide a good overlapping with the ligand (Schmidtke et al. JCIM 2010) but generally give no consensus in terms of pocket boundaries (Gao et al., Bioinfo 2013). In this work, we study the effects of the pocket estimation methods on the pocket-ligand pair classification. For that purpose, pockets of the 483 protein-ligand complex set (Pérot et al. PLoS One 2013) will be extracted using the four approaches. Each pocket-ligand pair set will be characterized using geometric and physico-chemical descriptors (Borrel et al. in submission) and will be classified according to its properties. The concordance of the four pocket-ligand pair classifications (where pockets are estimated using proximity, geometric, energetic, and template-based approaches) will be determined by different statistical criteria (Rand coefficient, percentage of concordant pairs) and will be analysed in terms of descriptors influence. These analyses will allow us to identify what are the effects of pocket estimation methods in the characterization and classification of pocket-ligand pairs, an important step for the pocket-ligand interaction prediction.

E59: Inès Rasolohery, Imen Daoud, Pierre Tufféry, Gautier Moroy and Frédéric Guyon. PatchSearch: a new method for surface patch comparison in proteins

Abstract: Specific recognition of a therapeutic molecule for its target protein has raised a lot of issues in the pharmaceutical field. Indeed, many promising therapeutic molecules could not be approved as drugs due to non-specific interactions which may lead to significant side effects.
Based on the structures of ligand-protein complexes, we defined an interaction patch like the residues involved in the interaction between a molecule and the complexed protein. Our study of the interaction specificity only focuses on proteins, since their conformation is known to be more rigid than ligands ones. In order to determine whether or not a ligand is able to interact specifically with a protein, we aim to assess the interaction patch preservation among other available protein structures in the PDB. For this purpose, we have developed a new method called PatchSearch, which compares protein surface patches.
Interaction patch and protein surfaces are extracted from complexes using NACCESS [1]. The matching between atoms composing the query patch and surfaces is performed by taking into

account the atoms nature and physicochemical properties of their residues to which they belong. Search for the patch onto the protein surface is obtained by the construction of a product graph. Product graph vertices are pairs composed of patch atoms which matched with surface atoms. Product graph edges link pairs with conserved neighborhood relationships. Patch search on a protein surface is based on a clique detection in the product graph [2]. However, this clique detection may be time-consuming. Therefore, in order to enable fast calculations, PatchSearch algorithm used quasi-cliques detection instead of cliques search [3, 4]. PatchSearch enables to recognize interaction patches with structural modifications. PatchSearch has been applied on protein-ligand benchmark [5].

1. Hubbard S.J. and Thornton J.M., Computer Program (1993)
2. Ullman J. R., J. Assoc. Comput. Mach (1976), 23, 31-42
3. Seidman et al. Social Networks (1983), 5(3):269-287
4. Batagelj and Zaversnik, A. D. A. C. (2002), 5(2):129-145
5. Kahraman et al., J. Mol. Biol. (2007), 283-301

E60: Sergei Grudinin and Georgy Derevyanko. HermiteFit: Fast fitting atomic structures into a low-resolution density map using 3D orthogonal Hermite functions

Abstract: Many algorithms in computational structural biology and structural bioinformatics deal with the exhaustive search in the six-dimensional space of translations and rotations of a rigid body. These algorithms are used, for example, in crystallography for molecular replacement, in computational biology to perform ligand docking, predict protein-protein interactions and discover structures of macromolecular assemblies. Modern exhaustive search algorithms either implement the fast 3D translational search using the fast Fourier transform (FFT) [1] or the fast 3D rotational search by means of the spherical harmonics decomposition and the FFT [2] or even the fast 5D rotational search [3]. Exhaustive search is also widely used as a preliminary step preceding the local search or flexible refinement procedures. Thus, the quality and the speed of the exhaustive search algorithms have a great impact on the solution of the vast variety of problems. Therefore, we believe that new directions of research on this topic are very important and highly valuable.

Here we present HermiteFit [4], a novel algorithm for fitting a protein structure into a low-resolution electron density map. The algorithm accelerates rotation of the Fourier image of the electron density by using 3D orthogonal Hermite functions. As a part of the new method, we present an algorithm for the rotation of the density in the Hermite basis and an algorithm for the conversion of the expansion coefficients into the Fourier basis. We implemented HermiteFit using the cross-correlation or the Laplacian-filtered cross-correlation as the fitting criterion. We demonstrate that in the Hermite basis, the Laplacian filter has a particularly simple form. To assess the quality of density encoding in the Hermite basis, we provide an analytical way to compute the crystallographic R-factor. Finally, we validate our algorithm using two examples and compare its efficiency with two widely used fitting methods, ADP_EM and colores from the Situs package, where our method outperform the two other in terms of speed per search point, while attaining the same level of accuracy.

The proposed algorithm can be straightforwardly applied to other problems in structural bioinformatics and computer science in general. HermiteFit will be made available at http://nano-d.inrialpes.fr/software/HermiteFit or upon request from the authors.

[1] Katchalski-Katzir et al. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. PNAS 89, 2195.
[2] J. Kovac & W. Wriggers (2002). Fast rotational matching. Acta Cryst D, 58, 1282.
[3] J. Kovac et al. (2003). Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. Acta Cryst D, 59, 1371.

[4] G. Derevyanko & S. Grudinin, HermiteFit: Fast fitting atomic structures into a low-resolution density map using 3D orthogonal Hermite functions. Acta Cryst D, accepted.

E61: Iain H. Moal and Juan Fernández-Recio. Training energy functions from changes in protein-protein binding affinity upon mutation

Abstract: The modelling of protein-protein interactions requires potential functions capable of quantifying the relationship between structure and binding free energy. Such potentials are frequently inferred from the statistics of structural databases and can be susceptible to systematic biases. An alternative approach is to infer intermolecular potentials directly from changes in binding free energy upon mutation. We develop this approach using ΔΔG values in the SKEMPI database. First, changes in intermolecular contacts between mutant and wild-type structures are calculated. Subsequently, potentials are fitted to the ΔΔG values via least-squares regression, using the assumption that ΔΔG can be approximated by the sum change in contact energies. The derived potentials are validated by their ability to rank docking poses and predicting absolute binding free energies.

E62: Isidro Cortes, Guillaume Bouvier, Michael Nilges, Luca Maragliano and Therese Malliavin. Enhanced conformational sampling of the catalytic domain of the adenyl cyclase CyaA from Bordetella pertussis

Abstract: The catalytic domain (AC) of the CyaA adenyl cyclase toxin of Bordetella pertussis undergoes a conformational transition from the inactive (calmodulin-free) to the active (calmodulin-bound) conformation upon binding to calmodulin. In the structure of the complex between AC and the C terminal lobe of calmodulin (C-CaM) elucidated by X-ray crystallography (Guo et al, EMBO J, 24:3190, 2005), AC displays an elongated shape. On the contrary, the calmodulin-free inactive conformation has been only qualitatively characterized as more globular by hydrodynamics measurements (Karts et al, Biochemistry, 49:318, 2010). To better understand at the molecular level the conformational transition of AC from the active to the inactive conformation, we introduce here a variant of the Temperature Accelerated Molecular Dynamics (TAMD, Maragliano and Vanden-Eijnden, Chem Phys Lett, 426:168, 2006), the soft-ratcheting TAMD (sr-TAMD). TAMD is a an enhanced sampling method designed to explore the free energy surface associated to a set of variables describing the process under study. In sr-TAMD, a soft-ratcheting criterion (Perilla et al, J Comput Chem, 32:196, 2011) IS introduced to accept values of collective variables proposed at each step of TAMD. Hence, the sr-TAMD allows us to drive the sampling of the AC conformational space to those regions where the protein displays a more globular shape. The resulting conformations were further clustered using Self-organizing Maps (Bouvier et al, Bioinformatics, 26:53,2010), allowing us to identify intra-protein hydrogen bonds specific of the appearance of the compact conformations.

E63: Thomas Coudrat, John Simms, Denise Wootten, Arthur Christopoulos and Patrick Sexton. Computer-aided drug discovery: development of a method for G protein-coupled receptor binding pocket refinement

Abstract: G Protein-Coupled Receptors (GPCRs) are a superfamily of transmembrane proteins that mediate cellular responses to their environment. Their activation is triggered by interaction of a ligand with their binding pocket at or near the extracellular face of the receptor that results in the recruitment of signalling effectors to the intracellular face of the receptor leading to the activation of signalling pathways inside the cell. With over 800 human GPCRs playing key roles in modulating tissue/cell physiology and homoeostasis, they represent a major target for pharmaceutical intervention and they are currently targeted by

over 30 % of marketed drugs.
In recent years major experimental developments in the crystallisation of membrane proteins have led to the determination of an increasing number of GPCR structures by X-ray crystallography. Computer-Aided Drug Discovery (CADD) methods are thus being increasingly applied to GPCRs leveraging the available structural information to rationalise drug discovery efforts. One of these methods used to find new drug leads is Virtual Screening (VS), which ranks libraries of small molecules based on the predicted complementarity between small molecules and the target GPCR binding pocket. The success of VS is highly dependent on the conformation of the binding pocket, therefore, a key step of CADD is to refine the binding pocket within a protein structure that is obtained from X-ray crystallography and/or homology models.
Here we present a new computationally efficient Ligand Directed Modelling (LDM) method for GPCR binding pocket refinement that aims to establish the global energy minimum of a GPCR binding pocket in complex with a known active ligand. Our LDM method samples the GPCR structure and docks a single known active ligand for that GPCR on each generated structure. All binding pocket/ligand complexes are scored and the best are selected to start a subsequent round of sampling, docking and selection. This results in a recursive refinement of the GPCR binding pocket in complex with its known active ligand.
To benchmark the method, we have used family A GPCR structures that have been crystallised with more than one known active ligand, and tested our LDM in a range of different refinement scenarios. In each LDM refinement experiment, the resulting structures were scored and compared to both the starting and final X-ray crystal structures using three metrics; (i) the binding pattern between the ligand and the binding pocket, (ii) the binding pocket shape and (iii) the binding pocket recovery of known active ligands in a small scale VS. This benchmark provides a guideline for the application of this LDM method in future CADD projects.

E64: Kliment Olechnovic and Ceslovas Venclovas. The CAD-score webserver: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes

Abstract: The Contact Area Difference score (CAD-score) web server provides a universal framework to compute and analyze discrepancies between different 3D structures of the same biological macromolecule or complex. The CAD-score method is based on calculating differences of contact areas derived from the Voronoi diagram of balls that correspond to heavy atoms of van der Waals radii. CAD-score was initially developed for proteins and recently has been extended to RNA/DNA 3D structures. The CAD-score web server accepts both single-subunit and multi-subunit structures and can handle all the major types of macromolecules (proteins, RNA, DNA and their complexes). It can perform numerical comparison of both structures and interfaces. In addition to entire structures and interfaces, the server can assess user-defined subsets. The CAD-score server performs both global and local numerical evaluations of structural differences between structures or interfaces. The results can be explored interactively using sortable tables of global scores, profiles of local errors, superimposed contact maps and 3D structure visualization. The web server could be used for tasks such as comparison of models with the native (reference) structure, comparison of X-ray structures of the same macromolecule obtained in different states (e.g. with and without a bound ligand), analysis of nuclear magnetic resonance (NMR) structural ensemble or structures obtained in the course of molecular dynamics simulation. The server is freely accessible at http://bioinformatics.ibt.lt/cad-score.

E65: Nadia Znassi and Andrey V. Kajava. Mapping amyloidogenicity to the known 3D structures of proteins

Abstract: Motivation: Various natively folded proteins being soluble contain aggregation-prone or amyloidogenic regions (ARs) in their amino acid sequences. The mechanisms by which these proteins avoid aggregation remain elusive.
Results: We used our program "ArchCandy" for the detection of ARs in the non-redundant set of proteins with known 3D structures. The analysis revealed that the predicted ARs are predominantly located within the α-helical regions. The observed tendency can be a mechanism by which proteins avoid aggregation since in order to form amyloid, the α-helical regions have to unfold and then refold into the β-structure. At the same time, the unfolded loop conformations of proteins usually have lower energetic barrier to fold in the amyloids and we found that the ARs are rare in the loop regions. Surprisingly, the ARs occur more frequently within the N- than the C-terminal regions.
Availability: ArchCandy was developed in Java and is available upon request.

E66: Claudia Caudai, Emanuele Salerno, Monica Zoppè and Anna Tonazzini. A multiscale model for 3D chromatin structure estimation using quaternions

Abstract: We present a method to reconstruct a set of plausible chromatin configurations from contact data obtained through Chromosome Conformation Capture techniques. We do not look for a unique configuration because the experimental data are not derived from a single cell, but from millions of cells. As opposed to most popular methods, we do not translate contact frequencies deterministically into distances, since this often produces structures that are not consistent with the Euclidean geometry. We build a data-fit function directly from the pairs of loci with the largest contact frequencies, assuming that they are likely to be in contact, and neglecting the pairs with very low or zero contact frequencies, as we cannot infer anything about their mutual distances. To obtain configurations consistent with both the data and the available biological knowledge, we introduce a chromatin model that can be suitably constrained. Taking advantage of the block structure of the contact matrix, we adopt a multiscale approach where the chromatin fiber is divided into a number of segments that can be treated in parallel. Each of them is modeled as a chain of partially penetrable beads whose properties (bead sizes, elasticity, curvature, etc.) can be constrained on the basis of biochemical and biological knowledge. The model parameters can easily be extended to exploit any further information available. Once the individual structures are reconstructed, each segment can be treated as an element of a new chain, and the procedure can be repeated recursively at different scales.
Our algorithm samples the solution space generated by the data-fit function through a Monte Carlo method. At each step, the subchains are perturbed by using quaternions. These are an extension of the complex algebra that offers a number of advantages, by avoiding singularities typical in the Euler matrix formalism, facilitating the composition of rotations, and allowing for a continuous evolution of the structure that is intrinsically compatible with the topological constraints.
To validate the new method, we applied it to real Hi-C data available online (Lieberman-Aiden et al., 2009). In particular, we analyzed the contact frequency data from the long arm of the human Chromosome 1 with a maximum resolution of 100 kb, obtaining a number of output configurations. For each configuration, the first division of the overall fiber included 25 topological domains (Dixon et al., 2012). The reconstructed structures were then assumed as single elements of a new chain (with nonuniform resolution), whose mutual interactions were estimated by the same algorithm. The output structures should be validated biologically. As a first test, we computed the relationships between the genomic and Euclidean distances of pairs of loci in the entire chains reconstructed. Our results are compatible with the analogous

plots, derived from FISH experiments on the same genomic region, found in Mateos-Langerak et al. (2009).

E67: Justas Dapkunas, Albertas Timinskas, Kliment Olechnovic, Mindaugas Margelevicius, Rytis Diciunas and Ceslovas Venclovas. The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures

Abstract: Easy-to-use computational tools for structural analysis of protein-protein interactions are valuable for studies of molecular mechanisms of biological processes. We have recently developed the PPI3D web server, which provides a possibility to search for experimentally determined 3D structures of protein complexes, to analyze the identified protein-protein interactions and to generate homology models of protein complexes. Structural data for experimentally determined protein-protein interactions are represented by the biological assemblies available from the Protein Data Bank. All the protein-protein interactions accessible through PPI3D are clustered according to both sequence and interaction interface similarity. This removes the redundancy of structural data while preserving alternative protein binding modes. The server enables users to explore interactions for individual proteins or interactions between a pair of proteins (protein groups). In both modes, structural data on protein-protein interactions are detected using sequence search with either BLAST or PSI-BLAST depending on the desired level of similarity. The PPI3D output enables users to interactively explore both the overall results and every detected interaction. The server reports the total protein-protein interface area, the sortable list of interface residues and their individual contribution to the interface area. The interface residues can be inspected in the sequence alignment or in the context of 3D structure. In addition, the server provides a possibility to construct a homology model for the protein complex. The server is freely accessible at http://www.ibt.lt/bioinformatics/ppi3d/

E68: Martino Bertoni, Marco Biasini, Florian Kiefer and Torsten Schwede. Comparative protein quaternary structure modelling using evolutionary interaction fingerprints

Abstract: Three-dimensional structures of protein provide valuable hints to describe its function. However, these macro-molecules rarely act alone and the large majority of soluble and membrane proteins are found in some aggregate form, with an amazing range of shapes and symmetries. With the increase in experimental information on protein quaternary structure and protein-protein interactions availability in public curated databases, structural modelling of such assemblies on a genome-wide scale has become a tangible goal.
During evolution, the quaternary structure of proteins is less conserved than their tertiary structure. This implies that even in the same protein family we can often observe a range of different oligomeric assemblies. For example, in the family of superoxide dismutase we find monomeric, dimeric and tetrameric forms. This observation obviously poses a challenge for modelling and prediction of protein structures. Several studies have been conducted to study properties of protein-protein interfaces, such as buried surface area size, energetic criteria, sequence conservation, amino acid composition, etc. However, so far no single criterion could be defined, which could reliably distinguish biologically relevant interfaces in protein models. Here, we introduce a novel descriptor of protein-protein interface conservation called “PPI fingerprint”. This descriptor is derived from the analysis of interface-surface entropy ratio in a series of multiple sequence alignments as a function of evolutionary distance. We observe that it reflects the distinct conservation patterns observed in various protein families, where the minimum value of PPI as a function of evolutionary distance reflects the optimum parameters at which the highest degree of interface conservation is expected.
In the context of protein structure homology modelling, we combine the new PPI fingerprint

measure with various classical interface descriptors using a machine learning approach based on random forests to assess protein-protein interfaces implied by different template structure complexes. The predictor has been cross-validated on a dataset of experimentally verified oligomeric structures. Here, we present the results of the validation, including the contributions of the various feature descriptors, and the overall performance of the predictor for modelling quaternary structures of proteins.

E69: Nolan Chatron, Florent Langenfeld, Virginie Lattard and Luba Tchertanov. In silico study of Vitamin K epoxide reductase and its mutants

Abstract: The vitamin K cycle occurs in the endoplasmic reticulum membrane. Several glutamic acids of selected coagulation factors are γ-carboxylated by the γ-glutamyl carboxylase, which oxidizes the vitamin K hydroquinone (VKH2) into the vitamin K epoxide (VKO) to generate functional clotting factors. The vitamin K epoxide reductase (VKOR) plays a key role in this cycle catalyzing the regeneration of VKH2 from VKO. VKOR reduces VKO into the vitamin K quinone (VK1) and further VK1 into VKH2. These reduction steps are catalyzed through an electron transfer mediated by four cystein residues (C43, C51, C132 and C135), from a protein partner (TMX, TMX4, ERp18) to the final electron acceptor, the vitamin K [1]. The structural data of the human VKOR are not currently available and their absence impedes to confirm the putative reaction [2], which supposed that a covalent intermediate is formed between a cystein residue of the active site C132XXC135 motif and the VKO carbon 3. VKOR is thus an important target in anti-coagulation therapies, but some patients develop VKOR's mutation non-sensitive to anti-vitamin K (AVK) inhibitors.
Our study of VKOR and its mutants aims to development new AVKs molecules overcoming resistance. Human VKOR in the native and mutated states was studied by in silico approaches. The effects of five different mutations (D36G, V54L, S56F, H68Y, Y139H) which influence VKOR activity and/or induce a resistance to AVKs, were examined by molecular dynamics (MD) simulations (VKOR was inserted into membrane, 200 ns trajectories). The MD data were used for characterisation of the mutation-induced effects on structure, dynamics and recognition properties of VKOR using Principle Component Analysis (PCA), Normal Modes Analysis, MDpockets [3] and Modular Network Analysis (MONETA)[4]. MONETA is based on concerted atoms motion along dynamics simulations, and brings out communication pathways (CP) between amino acids in the studied protein : it is an allosteric indicator. Thanks to MONETA, we highlighted some CPs which are affected by resistant mutations, indicating that one single residue mutation can disturd an other residue which is spatially distant in VKOR.
We identified a role of each studied mutation on structure, dynamics, the pockets profile and communication properties of VKOR. We were able to distinguished between the effects induced by activating mutations and resistant mutations. These data will be used for the development of mutant-sensitive inhibitors.
References
[1] Jin, D. Tie J. K., Stafford D. W. Biochemistry, 46:7279-7283, 2007.
[2] Davis C. H., Deerfield D.II; Wymore T.,.Stafford D. W, Pedersen L. G. J Mol Graph Model., 26:401-408, 2006.
[3] Schmidtke, P., Bidon-Chanal, A., Luque, F. J., & Barril, X. (2011). Bioinformatics. 27, 3276-3285.
[4] Allain, A., Chauvot de Beauchêne, I., Langenfeld, F., Guarracino, Y., Laine, E., & Tchertanov, L. (2014). Faraday Discussions DOI:10.1039/C4FD00024B., 1-18.

E70: Jessica Andreani and Johannes Soeding. Prediction of beta-strand interactions from direct coupling patterns

Abstract: The problem of protein structure prediction from sequence alone is one of the most important and difficult in computational biology. Recently, a major breakthrough was achieved in template-free protein structure prediction. When enough homologous sequences are available, correlated mutations can be detected using statistical analysis and contacts between pairs of residues can thus be predicted. Several methods from this class of approaches have been successful in predicting protein structures from sequence alone. So far, the main limitation of these approaches is the very large number of sequences that is necessary to extract signal from multiple sequence alignments.
The prediction of contacts between beta-strands is well suited for the use of predicted residue-residue couplings, even when those couplings are noisy. Indeed, interactions between beta-strands create patterns in the coupling matrices that can be detected more reliably than single couplings. Moreover, the prediction of beta-contacts is useful for protein tertiary structure prediction and it has been an area of active research for the past ten years.
We have developed an algorithm to predict the pairings of beta-strands, by detecting patterns in coupling matrices predicted from multiple sequence alignments using a state-of-the-art method. Our algorithm relies on statistical modeling and structure-guided detection of relevant patterns in the matrix of predicted couplings. It was tested on two datasets of protein chains containing beta-contacts, including cases with very few available homologous sequences. When the DSSP secondary structure state assignment is available, our method reaches similar recall and higher precision at the strand level, compared to previously published methods. In addition, our method displays considerably higher precision and recall at the residue level than previous methods and it outputs predictions for beta-bulges. Finally, unlike previously published beta-sheet contact prediction methods, our method can also make use of predicted secondary structure, thus dismissing the need for a structure-based assignment which is unavailable in reality when the protein structure is unknown.

E71: Kęstutis Timinskas and Česlovas Venclovas. Computational Analysis of DNA Polymerases and their Homologs in Bacterial Genomes

Abstract: DNA polymerases are essential to bacterial survival. All known bacterial DNA polymerases belong to five families: A, B, C, X and Y. Most experimental research has been done in Escherichia coli, which has polymerases of four families: Pol III (C family) - the primary polymerase in DNA replication; Pol I (A family) - a supporting replicative polymerase, also involved in DNA repair; Pol IV and Pol V (Y family) - translesion synthesis repair polymerases; Pol II (B family) - a repair and a putative backup replicative polymerase. Although biochemical data about polymerases from some other bacteria (Bacillus subtilis, Mycobacterium tuberculosis, Staphylococcus aureus) are available, it is insufficient to draw a broader picture.
To better understand the distribution, combinations and functions of DNA polymerases in bacteria we performed a large-scale analysis. We have collected DNA polymerases of all families from nearly 2000 fully sequenced bacterial genomes. It turned out that different polymerase families are easily separable from each other. We then analyzed each family in more detail: divided into smaller groups, identified known functional domains in all polymerase sequences and compared sequence and structure conservation between families and groups.
We found that polymerases of the C-family are encoded in all bacterial genomes without a single exception. Therefore, these polymerases appear to be universally responsible for chromosomal DNA replication in bacteria. Polymerases of A and Y families are also widespread among bacteria. Most of the A family polymerases cluster together according to the sequence similarity, suggesting that these polymerases in most bacteria may function similarly to Pol I in E. coli. On the other hand, polymerases of the Y family can be subdivided

into several functionally different groups: a relatively narrow group containing E. coli Pol V (easily discerned according to the unique C-terminal structural motif), several groups similar to Pol IV and two groups of ImuB (an inactive polymerase, involved in error-prone DNA repair). Interestingly, Pol V mediated SOS-response is not universal, but is not localized to few bacterial phyla either. ImuB is most often found together with DnaE2, a non-essential error-prone polymerase of the C-family. Interestingly, polymerases of the B-family are predominantly found only in Proteobacteria (beta, gamma and delta) and Chlorobi and are rare in other bacterial phyla. Apparently, the B-family, essential in eukaryotes, is not widespread among bacteria. Polymerases of the X family, absent in E. coli, were identified in about 25% of bacteria. Surprisingly, at least 2% of analyzed bacteria were found to contain only C-family polymerases in their genomes. This finding suggests that C-family polymerases alone may be sufficient to carry out all the essential tasks of DNA synthesis in the bacterial cell.

E72: You-Yu Lin, Mei-Ju May Chen and Chien-Yu Chen. A study of inter- and intra-protein correlated mutations on highly similar protein sequences

Abstract: Protein-protein interactions (PPIs) are essential for many biological processes. However, since PPIs with experimental validation are not sufficient in several species, many studies have focused on in silico prediction of PPIs. Some classical studies indicated that co-evolution of a protein pair is useful to identify PPIs since interacting proteins are probably mutated simultaneously under evolutionary selection. Considering co-evolution at residue level, correlated mutations can be classified into two groups. The first is about maintaining protein structure that usually happens in the core of proteins. The second is related to preserving the protein function, which are more frequently occurred on the surface of proteins. According to previous studies, functional constraints are stronger than structural ones, i.e., only a few residue combinations are admissible on functional residues. This study first showed that detection of co-evolution in multiple sequence alignment (MSA) is still possible on highly similar sequences. Next, we attempted to investigate the frequencies of inter- and intra-protein correlated mutations with a systematic way to evaluate which is a stronger constraint during evolution. Protein pairs of Homo sapiens (resolution of at least 3 Angstrom) were collected from Protein Data Bank (PDB). In this study, only heterodimers including two proteins were considered. To investigate correlated mutation of a PPI, the curated homologous sequence sets, including human and other 14 vertebrates, were collected from the Evola Database. By overlapping the lists of PPIs from PDB and Evola, a set of 1,153 physically interacting protein pairs with MSAs was constructed. MSA was applied on the homologous sets by using ClustalW2. For every two MSAs of a PPI, the correlated mutations were estimated by using McLachlan-based substitution correlation (McBASC). Finally, the identified correlated mutations were separated into two groups, inter correlated and intra-correlated by in-house scripts. We adopted the ratio: (Number of inter-protein correlated mutations) / (Number of inter-protein correlated mutations + Number of intra-protein correlated mutations) as the measurement to evaluate the strength of different co-evolution constraints. Using different correlation levels as the threshold, we observed that the resultant ratios are getting larger when the correlation threshold is higher, suggesting that the inter-protein correlated mutation is a stronger constraint than the intra-protein one. This study confirms the previous discovery that inter-correlated mutation is much stronger than intra, even on the highly similar sequences collected from Evola database. Besides, we observed that when using higher correlation threshold, the sites of correlated mutation seems to be more frequently co-occurred. It means those sites are highly related, and it makes sense that correlated mutation happened in cascading manner, not just in one place.

E73: Sucharita Dey and Bin Tean Teh. Molecular dynamic simulation reveals altered function of a chromatin regulatory protein involved in myeloid malignance due to single missense mutation

Abstract: Recent next generation sequencing efforts have revealed frequent mutations in epigenetic modulators causing a wide range of cancers. Of particular interest is the frequently mutated gene TET2 known to be involved in peripheral and angioimmunoblastic T cell lymphomas (PTCL, AITL)1,2. TET proteins are known to oxidize modified base - 5-methylcytosine (5mC) on DNA and play important roles in various biological processes. TET2 specifically recognizes CpG dinucleotide and shows substrate preference for 5mC in a CpG context. However, detail functional mechanisms and how mutations lead to cancer remain unclear. Owing to the recently solved TET2 crystal structure3, more detailed investigation on the structure-function relationship is made possible. The hotspot mutation position - Arg1261 found in AITL, for example, is found to interact with cosubstrate NOG (N-OxalyGlycine) and is critical for its activity; also it may have a role in distinguishing normal and modified base on DNA. Here, we examine the possible effect of this TET2 mutation on its structure and it's binding to the DNA substrate which may shed light on its roles in tumorigenesis.

Using molecular dynamics (MD) simulation, we found significant changes in the intra-atomic interactions within the protein. The mutation hindered the possible cation-pi stacking electrostatic interaction seen in the wild type where the modified DNA base (5mC) was sandwiched by an aromatic residue (Tyr1902) and the cation (Arg1261); this was hypothesized to have function in distinguishing modified DNA bases4. Although the total energy of both systems was very similar, the backbone fluctuation was greater in the mutant. The base stacking interaction between Tyr1902 of TET2 and the pyrimidine base of mc6 was observed in wild type TET2 throughout the time period of 50ns but was interrupted in the mutant as Tyr1902 was seen to move away from mc6; the architecture hence proved to be necessary for specific recognition of methyl-CpG dinucleotide by TET2. Also the distance between Arg/Pro1261 (Pro in mutant) and mc6 fluctuated significantly in the mutant. When checked for hydrogen bonding pattern throughout the simulation, 7 prominent h-bonds were found in the wild type among which 3 were involved with 5mC (modified base), which was completely lost in the mutant. The binding study showed decreased free energy of binding (delta G) between the protein and DNA in the mutant compared to the native; this further supported the possibility of loss of function in the mutant. Hence the MD simulation data suggested that the residue Arg1261 is crucial and its mutation to Pro might have an altered effect on 5mC-DNA substrate recognition and subsequent protein function.

[1] Palomero et al., Nature Genetics 2014;46:166-170.
[2] Yanagimoto et al., Nature Genetics 2013; 46:171-175.
[3] Hu et al., Cell 2013; 155:1545-1555.
[4] Tsai C & Tainer AJ. Cell 2013; 155:1448-1450.

E74: Alexander Monzon, Emidio Capriotti and Gustavo Parisi. Conformational diversity of protein functional regions improves the characterization of deleterious mutations

Abstract: The native state of the proteins shows a wide range of conformations, which are important for their biological function. The conformers characterizing the native ensemble show different degrees of complex biological activities, mingled in a dynamic equilibrium, define as a whole the structural basis of protein function. The flavor of these structural changes goes from the large relative movements of complete domains to the small changes in the rotation of residues side-chain. Altogether, these movements modulate the transit of ligands (substrate, products, ions, water and allosteric modulators) through pathways

connecting the surface of the protein with its interior. Tunnels, channels, cavities, pockets, grooves, voids and pores are some of the structural features of proteins defining the traffic of ligands inwards and outwards the protein accordingly with the different conformers in the native ensemble. Conformational diversity (CD) could produce variations in the size, wideness and deepness of these functional regions changing their physicochemical properties and then defining the differential biological activities observed in the conformers. Due to the biological importance of these regions we decided to study how deleterious-related mutations could occur preferentially associated with them. To this purpose we collected 382 proteins (3095 conformers) with 2394 mutations (1642 disease and 752 polymorphic) were each of the protein show experimentally probed CD extracted from CoDNaS database. Tunnels, cavities and pockets were estimated using Fpocket and MOLE programs. All the mutations were mapped into each of the conformers for each protein in the dataset to define to which functional region belong. We found that deleterious-related mutations occur preferentially in functional-regions (Fisher test p-value<0.005) in reference to the occurrence of polymorphic mutations. As it is well established that deleterious-related mutations involve mainly buried residues, using all the buried positions we test how deleterious-related and polymorphic mutations could be differentially associated with the functional regions. We found using a Fisher test that for buried residues the distributions of mutations are different with a p-value<0.005. Using the information of the CD of each protein, we found that deleterious-related mutations are less mobile that polymorphic mutations (KS test p-value<0.001). This trend is also observed when the deleterious-related mutations occurring in any of the functional-structures are compared with the polymorphic mutations also occurring in functional regions (KS test p-value<0.001 and Wilcoxon test p-value<0.01).
Our results indicate that the analysis of functional-regions such as tunnels, cavities and pockets and their CD can help to better understand the functional effect of protein mutations and contribute to the development of new bioinformatics tools to estimate their deleterious impact.

E75: Tomas Bastys, Nadezhda T. Doncheva, Hauke Walter, Rolf Kaiser, Mario Albrecht and Olga V. Kalinina. Molecular dynamics simulations reveal a potential reason for selecting of distinct mutation combinations in HIV-1 protease clinical isolates: a study of several clones with and without L76V resistance mutation

Abstract: HIV-1 protease L76V mutation has been found in clinical isolates from patients with therapy failure in some European countries. It is a major resistance mutation against the antiviral drugs fosamprenavir, darunavir, and lopinavir, but it has been observed to have a resensitizing effect against atazanavir, saquinavir, and tipranavir. This mutation substitutes an amino acid to a chemically similar one; and the corresponding amino acid position lies in the hydrophobic core of the protease. In this analysis, we have investigated the mechanisms of why this mutation has such a profound effect on protein susceptibility to drugs using methods of molecular mechanics and dynamics simulations.
Recombinant viral clones derived from three PI-resistant HIV-1 patients (FB15, GH9, and RU1) were generated with amplicons ligated into a matched HIV-1 deletion mutant. L76V was either reverted to wildtype or introduced into the protein sequence by PCR mutagenesis. Genotypic resistance testing was performed by Sanger sequencing, phenotypic resistance testing was performed by a recombinant virus assay (PRRT). 76 L and V protease variants in complexes with saquinavir and lopinavir were modeled using MODELLER [1] and then their dynamics was simulated in silico using GROMACS [2]. Interactions between drug and protease from resulting trajectories were analyzed using residue interaction networks tool RINalyzer [3].
Saquinavir acquires different conformations in different simulations of the protein

corresponding to the FB15 genotype with and without the L76V mutation, resulting in dissimilar residue interaction networks; whereas contacts of 76L/V are relatively stable. The conformation of the flaps and particularly the flap tips is unstable in the case of 76V, indicating a possibility of the flap opening. This genotype has wild type amino acids at positions M46 and V82. A mutation of these residues is favored by selection in clinical isolates [4, 5]. We hypothesize that instability of the closed flap structure could be the cause of these preferences.
Simulations of the proteins corresponding to RU1 genotype with lopinavir reveal a potential reason for differences in resistance with and without the L76V mutation: In the RU1(76L) case residues near the active site, 27G, 28A, 29D, and 30D, make more contact with the ligand than in the RU1(76V) variant. 76L influences these contacts through interactions with 30D.
No strong differences were observed in residue interaction networks for the GH9 genotype, indicating that simulation time has been probably too short.
We conclude that molecular dynamics simulations have proven to be promising approach to unravel molecular mechanisms underlying resistance of the HIV-1 protease towards its inhibitors.
[1] Sali et. al. J. Mol. Biol. (1993)
[2] Berendsen et al. Comp. Phys. Comm. (1995)
[3] Doncheva et al. Trends Biochem Sci. (2011)
[4] Knops et al. AIDS (2010)
[5] Charpentier et al. PLoS One (2013)

E76: Dominic Simm, Klas Hatje and Martin Kollmar. Waggawagga: a Web Service for the Comparative Visualization of Coiled-Coil Predictions and the Detection of Charged Single-Alpha-Helices (CSAHs)

Abstract: Coiled-coils belong to the most common structural motives for proteins. The sequences of coiled-coils are typically characterized by contiguous heptad repeats (a-b-c-d-e-f-g)n with hydrophobic residues in "a" and "d" positions and the remaining residues mainly charged thus favouring alpha-helical formation. In most cases, the individual alpha-helices are not very stable but become stabilized by wrapping around each other. Thus left-handed coiled-coils are formed, in which the hydrophobic residues are buried in the centre of the molecule. Coiled-coils can either be parallel, anti-parallel, homodimers, heterodimers, trimers, or any other oligomer. Several programs exist to predict coiled-coil regions like Marcoil, Multicoil or Paircoils. However, all these prediction programs are biased towards highly charged sequences. Charged residues are even found, although rarely, in "a" and "d" positions. Although predicted as homodimeric coiled-coils, many of these sequences indeed form stable single-alpha-helices instead. Here, we present a new web service for comparing and visualizing coiled-coil predictions and the distinction between coiled-coils and charged single-alpha-helices, which we called Waggawagga. The user can run Marcoil, Multicoil, Ncoils and Paircoils on the query sequence, and use Scorer 2.0, PrOCoil and LOGICOIL for oligomerisation prediction. The query sequence is visualized as helical wheel-diagram of parallel or anti-parallel homodimers, or parallel homotrimers, and as net diagram. With sliders the user can interactively move through the sequence automatically updating the visualizations and a discriminating score, which supports the detection of charged single-alpha-helices. Waggawagga is available at http://www.motorprotein.de/waggawagga

E77: Akito Taneda. A multi-objective genetic algorithm for multi-target RNA design

Abstract: Recently, synthetic RNAs have been paid much attention due to their potential as the functional molecules useful for controlling biological systems. To design the structural RNA sequences which have a desired function, we have to take secondary structure into account during the design process since RNA sequences often have characteristic secondary structures important for their functions. The aim of inverse folding is to automatically design an RNA sequence which folds into a user-prescribed secondary structure. Although various inverse folding algorithms for a single target RNA secondary structure have been proposed, the number of available inverse folding software (eg RNAdesign and Frnakenstein) for RNA multi-targets is still small. Since the inverse folding for multi-targets can be utilized to design the RNA sequences with metastable structures or conformational changes, it is useful for designing RNA devices in which a response to an input molecule is taken into account through a structural change.
In this presentation, I present a multi-objective genetic algorithm for the inverse folding with multi-targets, which is a new version of my previous genetic algorithm (MODENA) for the inverse folding of a single RNA secondary structure target. To extend my genetic algorithm to multiple target structures, I have developed genetic operators, mutations and a crossover, in which dependency graphs for nucleotide complementarity are taken into account. Dependency graph is the graph structure for representing the constraints on nucleotide assignment in the designed RNA sequence, as described in the paper of RNAdesign; to randomly generate compatible RNA sequences which can fold into multiple target secondary structures, such constraints among nucleotide positions are necessary. I will present benchmark results for my new multi-objective genetic algorithm for multi-targets and compare the performance with those of the previous inverse folding methods for multi-targets.

E78: Lukasz P. Kozlowski and Janusz M. Bujnicki. Identification of potential protelomerases and their target sites in publicly available genomes

Abstract: Protelomerase is a unique resolvase enzyme that is capable of recognition, cleavage and ligation of linear chromosomes or replicons with covalently closed hairpin termini (telomeres) (1). Most of known protelomerases were detected in pathogens; e.g., Borellia burgdorferi or phages. Due to the degeneracy of the recognition site of protelomerase, a high recombinant divergence of mobile elements is possible. This was speculated as one of the major forces shaping host-pathogen interactions. To date, only a few protelomerases with their target sites are known.
In this study, we have made a systematic screen of publicly available genomes (bacteria and phages) in order to identify new members of the protelomerase family based on sequence and structure homology of known members. First, out of more than 3,300 genomes stored in NCBI, we filtered out those which are linear. Next, all potential open reading frames were extracted (in all six frames). In order to find sequence homology, HHsearch was used (a HMM model was constructed based on the properties of known protelermerase sequences) (2). For the most promising candidates, homology models were built (3). These models were compared to known structures with respect to the conservation of the catalytic site and other functionally important parts. Moreover, to avoid potential false positives, we designed a special algorithm for the recognition of potential protelomerase target sites. These imperfect palindromic sequences (around 60-bp long) consist of three parts (H, S, and C) are attributed to recognition at three different domains of the protelomerase. Based on the target site of TelN from Escherischia coli phage N15 (4), a special set of rules was deduced. Only genomes containing both potential protelomerase homologue and the target site were considered in further studies.

E79: Marco Pietrosanto, Eugenio Mattei, Manuela Helmer-Citterich and Fabrizio Ferré. BEAM: A new method for RNA secondary structure motifs discovery

Abstract: Motivation.Functional regions of RNAs are often related to recurrent patterns in sequence and/or secondary structure called motifs, and they have been found to play an important role in RNA folding and interaction with other molecules. Of particular importance is the interaction with RNA-binding proteins (RBP), which is involved in the regulation of a large number of cellular processes. Among the available motif-finding tools, the majority focuses on sequence patterns, sometimes including secondary structure as additional constraints to improve their performance. Nonetheless, secondary structures may have their motifs too, and these motifs may be concurrent to their sequence counterparts or even be independent from them. During the last years some effort was put into research of structural motifs using advanced methods which require long pipelines or high computational efforts. Here we present a novel method for structural motif discovery taking advantage of a new encoding for RNA secondary structure named BEAR (Brand nEw Alphabet for RNAs). In BEAR, RNA secondary structures are described as a string encoded with a structural alphabet describing each RNA sub-structure (e.g. loops, interiors) and its size. This representation allows us to adapt methods developed for sequence analysis on BEAR-encoded secondary structures.
Methods.Secondary structure representation is a bottleneck in RNA structural motif discovery tools. The way in which the RNA is described affects both the algorithmic complexity and its accuracy. Dot-bracket notation is generally not suitable, and more complex descriptions are preferred, increasing computational time. Sequence motif discovery tools can exploit the powerful resources offered by string theory instead. Our approach for discovering structural motifs combines the knowledge of sequence motif-finding algorithms to BEAR encoding. Thanks to BEAR, dot-bracket can be converted into a string storing secondary structure information. This context-aware representation allows us to use approaches that are similar to known methods of motif discovery (such as MEME). In particular, BEAM uses simulated annealing to find the motif which maximizes the score, given as the information of a PSSM. MBR (Matrix of BEAR-encoded RNA) is used to correctly align sequences. MBR Substitution scores capture type and amount of structural variation that structurally similar, homologous, and/or functionally related RNAs can tolerate and their values were computed on alignments of BEAR strings obtained by encoding RNA alignments stored in Rfam database.
Results.We used Rfam database to test the ability of our method to recall known motifs as well as RNAcontext dataset of RBP-binding RNAs. Preliminary results show a promising ability of our method to identify known motifs in each of the datasets. Thanks to BEAM, not only we do find the same motifs reported in literature, but we are also able to retrieve a more precise structural context of the motif.

E80: Arumay Pal and Chandra S. Verma. Investigating molecular mechanism and dynamics of ErbB family ligand binding by molecular dynamics simulation

Abstract: The ErbB group of receptors are the members of tyrosine kinase (TK) family that play important role in multiple cellular processes. Among the four receptors in human system, except ErbB2, three members are known to be activated by specific growth factors (ligands) [1, 2]. Ligand binding to the extracellular domain causes the receptor dimerization and triggers intracellular kinase activity. However, over expression, mutation, or altered signalling of these receptors are linked to multiple human malignancies [3]. Over expression of ErbB is prevented by two classes of inhibitors: i) biologic inhibitors such as monoclonal antibodies that target the extracellular domain to block the binding of growth factors and hence prevent

signal transduction, and ii) small molecules that compete with ATP to bind intracellular TK domain and suppress activity. Most of the existing biologic inhibitors are receptor specific, and some of them turn out to be ineffective due to certain mutations in the receptors. Understanding the molecular determinants for specific ligand binding are crucial and may provide insight into the canonical properties of different ligands in terms of their conformation and energy [4], which will be valuable in designing improved biologics.
In this study we used molecular modeling, canonical molecular dynamics simulations and free energy calculations (MM-GBSA method) to investigate the interaction dynamics of two high affinity ligands with their respective receptors— 'epidermal growth factor (EGF)' with ErbB1 and 'neuregulin (NRG)' with ErbB4, modeled from available crystal structures. Six other high to moderate affinity ligand-receptor complexes (three each for ErbB1 and ErbB4) were modeled and used as control dataset. The goal was to characterize how the structure and energy correlate with specificity of ligands to their respective binding partners. Comparison between energetic footprints representing van der Waals and Coulombic per-residue contributions showed conserved as well as variable residues are important for binding. For both EGF and NRG, stable hydrogen bonds were found at the B-loop and C-loop regions to clutch them between the domain I and III of the receptors. Receptor-ligand distances were similar in both the cases suggesting similar dynamic behaviour. Furthermore, clustering of ensemble of structures obtained from simulations can highlight the possible overlapping conformations and energetics of these growth factors. Thus the knowledge of structural information observed in binding may be applied in designing bi-specific peptide mimics that can target both ErbB1 and ErbB4; moreover therapeutics can be designed to target the ligands itself.
[1] Yarden and Sliwkowski (2001) Nat Rev Mol Cell Biol. 2:127-137.
[2] Endres et al. (2013) Cell 152:543-556
[3] Hynes and MacDonald (2009) Curr Opin Cell Biol. 21:177-184.
[4] Pan et al. (2013) Drug Discov Today 18:667-673

E81: Nadezhda T. Doncheva, John H. Morris, Olga Voitenko, Tomas Bastys, Karsten Klein, Eric F. Pettersen, Dina Schneidman, Andrej Sali, Thomas E. Ferrin, Olga V. Kalinina and Mario Albrecht. Analyzing the dynamic nature of proteins using residue interaction networks

Abstract: In the last years, a new interdisciplinary area of research that combines network science and structural biology in the context of visual analytics has emerged. By representing protein structures as networks of interacting residues, we can facilitate the study of structure-function relationships and gain more insight into complex molecular mechanisms such as protein-protein and protein-ligand interactions. Recently, we introduced a software suite that supports interactive, multi-layered visual analysis of protein structures and their interactions involved in protein binding, allostery, drug resistance and other molecular phenomena. In particular, we demonstrated how our integrative approach can be applied to visually analyze the impact of protein sequence mutations on its structure and function.
The current approach is, however, still limited to single static snapshots of proteins and their interactions. To capture the dynamic nature of protein structures and interactions, we developed a new method for visualizing and analyzing ensembles of protein structures. We use dynamic, weighted residue interaction networks that are capable of capturing the different protein conformations within the ensemble. We also facilitate the comparison of two ensembles within one network by highlighting the most similar and dissimilar residue interactions as well the rate at which they are present in the ensembles. In addition, we provide automatic enrichment of residue interaction networks with external data such as evolutionary conservation and physico-chemical residue properties to support the interpretation of interaction differences, in particular, upon amino acid changes.

As a proof of concept, we demonstrated how this approach can be applied on data from molecular dynamics simulations to characterize sequence mutations. Furthermore, we perform a thorough analysis of ensembles of docking structures (decoys) to evaluate their quality and aid in identifying the most probable docking interfaces.

E82: Peter Cimermancic, Patrick Weinkam, Justin Retternmaier, Daniel A. Keedy, Rahel Woldeyes, James A. Wells, James S. Fraser and Andrej Sali. Expanding the druggable proteome by characterization and prediction of cryptic binding sites

Abstract: Many proteins have small molecule-binding pockets that are not easily detectable in the ligand-free structure. These cryptic sites require a conformational change to become apparent. Therefore, understanding and accurately identifying these sites would expand the scope of ligand and drug discovery. Currently, cryptic sites can be identified experimentally by site-directed small-molecule tethering and computationally by long molecular dynamics simulations. Here, we begin by constructing a set of structurally defined apo-holo pairs with cryptic sites. Next, we comprehensively characterize the cryptic sites in terms of their sequence, structure, and dynamics attributes. We find that a cryptic site tends to be as conserved in evolution as a traditional binding pocket, but is less hydrophobic and more flexible. Relying on this characterization, we also use machine learning to predict cryptic sites with relatively high accuracy (for our benchmark, the true positive and false positive rates of 73% and 29%, respectively). Finally, we predicted cryptic sites in the entire structurally characterized human proteome (11,201 structures, covering 23% of all residues in the proteome). The method increases the size of the potentially "druggable" human proteome from estimated ~40% to ~78% of disease-associated proteins.

E83: Pedro Rafael Costa and Ney Lemke. A heuristic approach to study the influence of transcription pausing on RNA folding

Abstract: Uncovering the native and active structures of functional noncoding RNAs is imperative to understand the cell machinery, but predicting the folding pathway for the final tertiary structure from primary sequence remains a challenge for biophysicists and computational biologists. In this work we propose a heuristic approach to study the effects of transcriptional pausing on RNA folding. We considered that the RNA starts to fold as soon as it is synthesized, so, beside the sequence itself, the the user must provide a list with the dwell times for each nucleotide incorporation. Essentially, our code tests if there is enough time to fold the newly transcribed RNA between two subsequent incorporations and, if so, it returns a set with all the new possible topologies resulting from adding or removing a single helix from the current structure. Then, it assigns for each one a transition rate based on the difference between their free energy and the free energy of the current conformation using the Boltzmann weight. To determine these free energies, we used the parameters from Turner group, but also implemented the Isambert group approach to include pseudoknots. After that, the software performs a Monte Carlo simulation to set which structure will be set as the new conformation of the nascente RNA molecule and it incorporates a new nucleotide to the RNA chain, repeating these steps until the stopping criterion is achieved. The user can also include structure constrains, i.e., the base pairs that must be in the final structure, improving the accuracy of the simulation. To keep the results strictly computer-based, we determine the dwell times using the sequence-dependent model for transcriptional elongation from Wang group. Our code was entirely written in Wolfram Mathematica language and returns a set of potentials suboptimal structures, each one with its relative proportion. We benchmarked our method on 1254 structures from the RNAstrand database, covering from 9 to 350 nt long. About 90% of the dataset structures do not have pseudoknots, but we did not consider this a

priori. As preliminary results, we achieved 38% of sensitivity and 42% of specificity considering just the most frequent structure in each simulation. We intend to make the software available as a webservice in future.
Supported by: FAPESP (2012/19377-4), CNPQ (152838/2012-0).

E84: Aya Narunsky, Haim Ashkenazy, Rachel Kolodny and Nir Ben-Tal. Using PDB to explore conformational space of query proteins with at least one known conformation

Abstract: Proteins often alternate between several conformations as an important part of their function, e.g., active and inactive states of receptors, open and closed states of channels, etc. The Protein Data Bank (PDB) is a key tool in studying the conformational changes of a protein as in many cases, more than one structure can be found for the same protein. By looking at a large collection of conformational changes, it is possible to test various hypotheses regarding the behavior of the conformational space of proteins. We studied the conformational space of a non-redundant set compromised of 30,000 proteins. We discovered that in the majority of cases, a protein has more than one structure available in the protein databank. Also, in many cases two different proteins sharing structural similarity in one conformation share additional conformations as well.
Very often only one conformation of a protein is known, and the prediction of additional (biologically-relevant) conformations of a protein can provide more insight into its function in health and disease. We developed a method and a web-server aiming at predicting such possible conformations. Our method, ConTemplate, suggests an ensemble of conformations for a query protein with at least one known conformation, based on its similarity to other proteins in the PDB, and alternative conformations of these proteins.
Briefly, ConTemplate creates an ensemble of conformations for the query using the following three steps process. First, the entire PDB is scanned, and proteins whose structural-similarity to the query is above a preset threshold are collected. Second, for each of the collected proteins, additional known conformations are indicated, and clustered. In the third step, the server calculates models of the query in various conformations using the structure-based sequence alignments found in the first step, and the centers of the clusters found in the second step as templates.
We demonstrate the method with the kinase domain of the Epidermal Growth Factor Receptor (EGFR). Using the inactive conformation as our query, we reproduce the active conformation with root mean square deviation (RMSD) of 1.76Å, based on the query's structural similarity to the inactive conformation of Abl tyrosine-kinase, together with the known active conformation of the latter kinase. The sequence identity between the two kinase domains is only 40%, and the fact that they share similar active and inactive conformations might not be obvious.
The idea of inferring new conformations of a protein of interest based on known conformations in related proteins is not new. However, to the best of our knowledge, ConTemplate is the first automated implementation of this approach.

E85: Katerina Taškova, Marco Carnini, Sonika Rao and Andreas Hildebrandt. Exploring the space of co-optimal alignments for template-based protein model quality assessment

Abstract: With the increasing number of experimental structures in the Protein Data Bank (PDB), more protein structure are available as potential templates in Template-Based Modeling (TBM) of (target) proteins with unknown 3D structure. The quality of 3D structures predicted with TBM relies on good template choices as well as target-template alignments. Several studies in the literature, assessing the 3D structure prediction quality, point out that sub-optimal alignments can lead to predictions that are biologically more relevant than those

obtained with the optimal alignment. The latter especially matters when we need to model proteins for which a good template cannot be found in the PDB database. Such target-template pairs lead to a large set of unique co-optimal pairwise global alignments (having best alignment score), with potentially different impact on the model quality. In this context, we performed a large TBM experiment that allows us to study the model quality as a function of the alignment quality.
The experimental analysis involved proteins for which a 3D structure is experimentally available, selected according to the SCOP2 classification. More precisely, we selected proteins belonging to different types (globular, membrane, fibrous and unstructured), different folds (e.g., Globin-like proteins) and different secondary structures (mainly alpha, mainly beta, alpha or beta, alpha and beta alternated). We considered non-redundant sequences with length between 50 to 4000 and classified them based on the number of chains (single or multiple), number of domains (single or multiple) and resolutions (less than 2.5, less than 3, and greater than 3). Furthermore, we included multiple chains for protein complexes and single chain for the other proteins. The selected proteins were modeled using the MODELLER tool for comparative TBM, with a unique set of alternative pairwise alignments corresponding to a different combination of alignment parameters (gap opening/extension and scoring matrix). While we selected one protein per fold as a target for modeling, we chose few template proteins that are more/less significantly similar to the target sequence (with the MODELLER's function for profile building) to further analyze the model quality as a function of both template and alignment quality.
We analyzed and validated the quality of the obtained models with different quality scores, such as DOPE, RMSD, TM-score, GDT, and CMO. Their selection was driven by the intent to optimize different aspects of the model quality as well as to better capture the influence of the co-optimal alignments on the the model quality. Finally, we store all the data regarding the modeling process, including the target/template proteins, the models, and the calculated quality scores in an open access database for easy and efficient manipulation and analysis of the alignment quality influence.

E86: Yves Dehouck and Alexander S. Mikhailov. A sequence-specific elastic network model for coarse-grained studies of protein dynamics

Abstract: To understand how a protein achieves its biological purpose, it is quite often necessary to unravel the complexity of its dynamical behavior. Highly detailed approaches such as molecular dynamics are commonly used to tackle this issue, but considerable efforts have also been devoted to the development of more coarse-grained descriptions, which allow to follow protein motions on larger timescales, and to gain a better understanding of the general principles that govern the dynamical properties of proteins. In the elastic network model (ENM), residues are represented as single particles and connected to their neighbors by Hookean springs. Within this framework, the conformational fluctuations around a given reference structure can be studied with very reasonable computational costs. This elegant simplicity certainly contributed to the popularity of the ENM, and it has been successfully exploited in a wide range of applications. Yet, despite its many achievements, the ENM also comes with a severe limitation, as it has so far been unable to account for the chemical specificities of the different amino acids types. We present a method to derive a sequence-dependent force field for the ENM, from the statistical analysis of an extensive dataset of NMR conformational ensembles. This new force field is shown to yield a strongly improved description of the cooperative aspects of residue motions, demonstrating that the ENM does not have to be exclusively structural, and that sequence details may be allowed to play a major role in coarse-grained descriptions of protein dynamics. Thereby, our study paves the way towards systematic investigations of the effect of mutations on protein dynamics and, more

generally, comparative analyses of motions in proteins that share a similar structure but present differences in sequence.

E87: Marco Pasi, John Maddocks, Richard Lavery and The Ascona B-Dna Consortium. Microsecond-scale sequence-dependence of B-DNA mechanics and cation binding.

Abstract: DNA carries out its function in the cell by interacting with proteins. Proteins bind to DNA both by "reading" its sequence and by probing its deformability. The sequence dependence of the mechanical properties of DNA is at the basis of these "indirect recognition" processes. Understanding this second layer of genetic information requires extensive structural and dynamical knowledge of B-DNA, which is currently unavailable from experiment.
An international group of laboratories (The Ascona B-DNA Consortium, or ABC) got together in 2002 with the aim of using molecular simulations to obtain such information. The ABC laboratories' joint effort allowed the collection of more than 60 microseconds of molecular dynamics (MD) simulations of a specifically designed set of DNA sequences in water at physiological salt concentration. The resulting database is the most complete collection of simulations for the systematic study of the molecular-detail mechanical properties of B-DNA.
Detailed analysis of the trajectories allowed us to identify local B-DNA conformational substates, to exhaustively assess the sequence-dependent variations in the relative population of the substates and link these variations to the formation of specific interactions within the double helix. The surprisingly simple patterns that arise clarify the mechanisms that lead to multiple substates and allow us to determine in which cases substates will arise.
Furthermore, thanks to the recently developed curvilinear helicoidal coordinate analysis of ion distributions around DNA, we were able to shed light on the sequence-dependent variability of cation densities, and how this variability couples to the conformational fluctuations of the double helix.
These results represent an important step towards a detailed mechanical understanding of base-sequence effects on B-DNA, and constitute a valuable resource for developing coarse-grain models of DNA.
** The Ascona B-DNA Consortium is composed of the Authors and David Beveridge, Thomas C. Bishop, David A. Case, Thomas Cheatham III, Jeremy Curuksu, B. Jayaram, Filip Lankas, Charles Laughton, Roman Osman, Modesto Orozco, Nada Spackova, Jiri Sponer and Krystyna Zakrzewska.

E88: Juergen Haas, Alessandro Barbato, Steven Roth, Tobias Schmidt, Konstantin Arnold, Khaled Mostaguir, Lorenza Bordoli and Torsten Schwede. The ProteinModelPortal - How good is my modeling ? First Results From The Continuous Automated Model EvaluatiOn (CAMEO)

Abstract:
Protein structure modeling is widely used in the life science community to build models for proteins, where no experimental structures are available. However, depending on both the specific target protein and the applied modeling approach, the accuracy of the structural models may vary significantly between different modeling servers. The ProteinModelPortal (PMP, http://www.proteinmodelportal.org) contains more than 21 million models for more than 5.1 million distinct UniProt sequences from many well-known modeling servers such as ModBase(1) and SWISS-MODEL(2) repository as well as specialty modeling resources such as GPCRDB(3) and is regularly updated. PMP(4) uniquely allows users to search both for experimental structures deposited to the PDB (5) and the pre-computed models from the

modeling partners. A first indication of the expected model quality is displayed on PMP, but standards for assessing model quality are still being formulated by the community. The critical assessment of protein structure prediction (CASP) has been reviewing participating modeling servers every two years(6) with a selected set of targets. The Continuous Automated Model EvaluatiOn platform(CAMEO, http://www.cameo3d.org), however, is weekly assessing accuracy and reliability of the servers, thereby collecting a high number of targets. CAMEO, was inspired by two previously developed, but deceased evaluation systems, which were still post-PDB release based(7).
The CAMEO platform contributes to the quality standard discussion by continuously assessing the performance of servers predicting protein structures (3D), ligand binding site residues (LB) and quality of models (QE). Emphasis in CAMEO across all categories is put on evaluating many scores reflecting the different aspects of e.g. for fitting EM density maps, where coverage might be favored over accuracy. Continuous assessment of prediction servers allows to retrospectively analyze the performance of a given server, allowing to select the top performing servers for a given task. CAMEO supports the developers of prediction servers, as they can straightforwardly test new developments and monitor the performance of the productive servers continuously.
References
(1) U. Pieper, B. Webb, G.Q. Dong, et.al. Nucleic Acids Res. 42, 336-346, (2014)
(2) F. Kiefer, K. Arnold , M. Künzli Nucleic Acids Res. 37, D387-D392 (2009)
(3) V. Isberg, B. Vroling, R. van der Kant, et.al. Nucleic Acids Res. (2014) Jan;42(Database issue):D422-5
(4) J. Haas, S. Roth, K. Arnold, et.al. Database bat031 (2013)
(5) P.W. Rose, C. Bi, W.F. Bluhm, et.al. Nucleic Acids Res. (2013) Jan;41
(6) a) K. Joo, J. Lee, S. Sim, et.al. Proteins. (2013) doi: 10.1002/prot.2439; b) V. Mariani, F. Kiefer, T. Schmidt, J. Haas, T. Schwede Proteins 79 Suppl 10:37-58 (2011)
(7) a) L. Rychlewski, D. Fischer Protein Sci 14(1):240-5 (2005); b) V.A. Eyrich, M.A. Martí-Renom, D. Przybylski,et.al. Bioinformatics 17(12):1242-3 (2001)

E89: Qingzhen Hou, Jaap Heringa and K. Anton Feenstra. Differential Conservation Between Interacting and Non-interacting Homologs Identifies Interface Residues

Abstract:
Many protein families participating in protein-protein interactions have several sub-families that bind to different partners or sub-families that do not interact. Specificity in these interactions is often decisive to functions of proteins involved in the interaction[3]. Therefore, in addition to conservation that is typically used by methods to predict interface sites from protein sequences, the specificity between the interacting versus non-interacting groups might be used for recognising the interaction sites.
All homodimers and monomers from the PISA database are collected in a local database, which is then used to BLAST All-against-All. A homodimer group is formed by all homodimer hits with e-values lower than the first monomer hit; likewise for monomers. A homodimer-monomer pair group is created by matching a homodimer group with the its first monomer hit, and its corresponding monomer group. The alignment of the pair group is used for further analysis and prediction of interface sites. Validation of predictions is done based on the surface and interface site definitions in the PISA database.
Our results show that in this dataset, the SH [1]signal is able to distinguish interface and other surface residues. Furthermore, we find limited overlap with interface positions predicted by two other sequence-only methods: ISIS [4] and SPPIDER [5]. We have shown it is possible to predict interaction sites out of all residues using nothing more than sequence and group

specificity information. Further improvement of prediction accuracy could be expected from including additional features of sequences, such as neighbour support information. Moreover, our analysis will be helpful to gain specific insights into specificity and conservation evolutionary selection mechanisms of protein-protein interactions.

[1] Feenstra, KA, W Pirovano, K Krab & J Heringa. Sequence Harmony: Detecting Functional Specificity from Alignments, Nucl. Acid. Res. 35 W495 2007 www.ibi.vu.nl/programs/seqharmwww
[2] Pirovano, W, KA Feenstra & J Heringa Sequence comparison by sequence harmony identifies subtype specific functional sites. Nucl. Acids Res. 34 6540–6548 2006
[3] Feenstra, KA, G Bastianelli & J Heringa. Predicting Protein Interactions from Functional Specificity. in: From Computational Biophysics to Systems Biology – NIC Series 40 89-92 2008
[4] Porollo A, Meller J. Prediction-based fingerprints of protein–protein interactions[J]. PROTEINS: Structure, Function, and Bioinformatics, 2007, 66(3): 630-645.
[5] Ofran Y, Rost B. ISIS: interaction sites identified from sequence[J]. Bioinformatics, 2007, 23(2): e13-e16.

E90: Qingzhen Hou, Kamil Krystian Belau, Marc F. Lensink and K. Anton Feenstra. Mapping the Protein-protein Interaction Free Energy Landscape

Abstract: Protein-protein complexes are involved in most biological processes. In order to understand how protein complexes are formed, a thorough knowledge of the process of binding is critical. Although many efforts have been devoted to the development of methodology for this purpose, prediction from both sequence as well as protein docking methods have their respective limitations.

In a previous study, we have successfully used a coarse-grained forcefield, MARTINI, to calculate the potential of mean force (PMF) of the dissociation of two interacting proteins, starting from the crystal-structure of the complex.(May et al., 2014, Bioinformatics 30:326).Compared to atomistic methods, the speed-up by coarse-graining is about 500-fold. Here, we show that our method also performs well on other proteins, and that were not in the initial test set. Furthermore, we show how the PMF compares between native binding and nonspecific binding. To this purpose, we have obtained a blind dataset from the ensemble of CAPRI (Critical Assessment of PRediction of Interactions; Lensink & Wodak 2013, Proteins 81:2082) experiments. This dataset contains 15 target protein complexes which were submitted to the experiment by expert docking groups. For each target, around 1000 predicted bound conformations are present in the CAPRI dataset. For each of these we calculated the PMF of unbinding from which we obtained a reverse estimate of the free energy of binding in that orientation. In most cases, we are able to identify native-like binding orientations from non-native ones from the calculated free energies. Moreover, the methodology enables us to map these free energies onto a free energy landscape of binding. This allows us to investigate the underlying physical mechanism for protein-protein binding, in detail and at an unprecedented scale.

(1) Kastritis, P. and Bonvin, A. (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J. Proteome Res., 9(5), 2216C2225.
(2) May, A., Pool, R., Van Dijk, E., Abeln, S., Heringa, J. and Feenstra, K.A. (2013) Coarse-grained simulations: realistic binding free energies for real proteins realistic binding free energies for real proteins. Bioinformatics 30 (3), 326-334
(3) Lensink, M. F., & Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. Proteins: Structure, Function, and Bioinformatics, 81(12), 2082-2095.

E91: K. Anton Feenstra, Tom L.G.M. van den Kerkhof and Esther F. Gijsbers. Application Showcases for Sequence Harmony: Specificity Detection in HIV Protein Sequences

Abstract: Motivation: HIV evolution in the patient is key to finding effective treatment options, but the mechanisms at work are only partially understood. Here, present the application of our Sequence Harmony method, a simple entropy-based method to detect differences in amino acid compositions between groups of related protein sequences, as implemented in the multi-Harmony webserver. In the capsid protein, we compare HIV capsid sequences from progressing and from non-progressing (HLA-B57 immunotype) patients, in order to identify immunotype dependant 'escape' mutations and subsequent compensatory mutations. In the envelope protein, we compare HIV-1 clonal variants from 31 individuals with diverse levels of CrNA 2–4 years post-seroconversion to find unique characteristics that support the induction of cross-reactive neutralizing activity that occurs in some individuals.
Results: In capsid, we find putative escape mutations, which are a trade-off between evading immune pressure and maintaining viability; a few of these sites reside in known B57 epitopes. Additional mutations, infer improved replication rates, consistent with subsequent compensatory mutations. In envelope, we identify a number of amino acid changes associated with the development of CrNA. These residues mapped to various Env subdomains, but in particular to the first and fourth variable region as well as the underlying α2 helix of the third constant region.
Conclusion: We show that, using the multi-Harmony specificity detection tool, we can identify known HLA-B57 epitopes as well as previously unknown sites in the HIV capsid protein that, when mutated, significantly influence in-vitro replication rates. In addition, we can identify and refine regions that associate with the induction of CrNA in the HIV envelope glycoprotein. This highlights the usefulness of Sequence Harmony for analysing large sequence datasets, allowing experimental efforts to be focused on the most promising regions.
1) E.F. Gijsbers, K.A. Feenstra, A.C. van Nuenen, M. Navis, J. Heringa, H. Schuitemaker & N.A. Kootstra. PLoS ONE 8:12 (2013) doi:10.1371/journal.pone.0081235.
2) T.L.G.M. van den Kerkhof, K.A. Feenstra, Z. Euler, M.J. van Gils, L.W.E. Rijsdijk, B.D. Boeser-Nunnink, J. Heringa, H. Schuitemaker & R.W. Sanders. Retrovirology 10:102 (2013) doi:10.1186/1742-4690-10-102.
3) B.W. Brandt, K.A. Feenstra & J. Heringa, Nucl. Acids Res., doi:10.1093/nar/gkq415 (2010).

E92: Irene Farabella, Daven Vasishtan, Agnel-Praveen Joseph, Arun Prasad Pandurangan and Maya Topf. Validation of 3D Electron Microscopy Density Fits assembly models using TEMPY

Abstract: The fitting of atomic structures into 3D electron microscopy density maps is now routinely applied to gain further insight into the macromolecular assemblies they represent. Most fitting programs use the cross-correlation coefficient to estimate the goodness-of-fit between a given atomic model and a density map or between two maps. To complement this, other scoring methods have been proposed, such as a mutual information score and a normal vector score [1]. However, different scores will have different advantages depending on the scenario and currently there are no tools that allow the assessment of fit quality using a large selection of scoring methods in a single platform [1,2]. Here, we implement such a platform called TEMPy. TEMPy is a Python library which provides a set of functionalities for atomic model and density map processing as well as for fit validation using a variety of scoring functions.
1. Vasishtan D, Topf M (2011) Scoring functions for cryoEM density fitting, Journal of structural biology,174: 333-343.

2. Henderson, R., et al. (2012) Outcome of the first electron microscopy validation task force meeting, Structure, 20, 205-214.