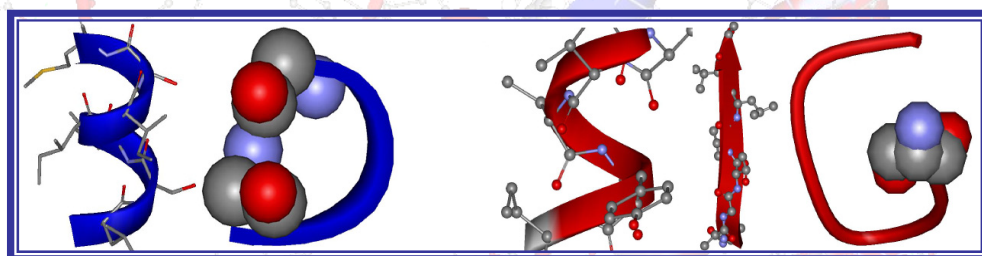




The 1st Structural Bioinformatics Meeting at ISMB/ECCB2004



29-30 July 2004, Glasgow, Scotland, UK



<http://3Dsig.weizmann.ac.il>

Table of Contents

	Page
Sponsors	3
Committees	5
Program	7
Abstracts	
Invited Speaker Abstracts	9
Speaker Abstracts	12
Laptop Session Abstracts	30
Abstract Index	97
List of Authors	107
List of Participants	109

Roche – **3Dsig** Gold Sponsor:



Any **PLANS** for
Start Your Career with Roche.
Tomorrow?
www.careers.roche.ch

PLANS
PLANS for Tomorrow.

*At Roche we contribute to improving people's health and quality of life by developing and marketing innovative therapeutic and diagnostic products and services. Your ideas could help shape tomorrow's innovations in healthcare. Plans are life's roadmap to the future. Come realise your plans with us:
www.careers.roche.ch*



3Dsig Sponsors:

Coming soon to the UK...

Biosystems Informatics Institute (BII)

Specialising in analysis and modelling of biological complexity

Recruiting IT specialists (all levels) in:

- proteomics
- constraint-based dynamic systems modelling
- protein-protein interaction prediction, eg CAPRI
- *ab initio* protein structure determination
- bioinformatics software engineers

We are eager to see your CV:

Biosystems Informatics Institute
2nd Floor, Bioscience Centre
International Centre for Life
Times Square
Newcastle upon Tyne NE1 4EP

tel : 0191 211 2560
fax : 0191 211 2561
web: www.ifbproject.info
e-mail: enquiry@ifbproject.info



Biosystems Informatics Institute

The Biosystems Informatics Institute receives funding from:  



מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

3Dsig Committees

Organizing Committee:

- Ilan Samish, Weizmann Institute of Science, Israel
- Marvin (Meir) Edelman, Weizmann Institute of Science, Israel

Scientific Committee:

- John Moulton - Chair, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, USA
- Phil Bourne, University of California San Diego, USA
- Luis Serrano, European Molecular Biology Laboratory, Heidelberg, Germany
- Michael Sternberg, Department of Biological Sciences, Imperial College, London, UK
- Ron Unger, Faculty of Life Sciences, Bar-Ilan University, Israel
- David Gilbert (Honorary Member), Department of Computing Science, University of Glasgow, Scotland, UK
- Janet Thornton (Honorary Member), European Bioinformatics Institute, UK

3Dsig Program for Day 1, Thursday, July 29th 2004

8:45 *Opening remarks* – **Marvin (Meir) Edelman & Ilan Samish**

Docking and Assemblies (Chair - David Gilbert), 9:00 – 10:15

9:00 **Rob Russell** (invited presentation): A Structural Perspective on Protein Interactions & Complexes. (Invited-1)

9:30 **Maxim Tortov**: Protein-Protein Docking Simulations with Local Backbone Flexibility. (36)

9:45 **Q. Jane Su**, Orit Foord, Scott Klakamp, Holly Tao, Larry L. Green: Modeling and Simulation of Antibody-Antigen Interaction: An Integrated Approach. (72)

10:00 **Yuval Inbar**, Haim J. Wolfson, Ruth Nussinov: Prediction of Multi-Molecular Assemblies by Multiple Combinatorial Docking. (76)

10:15 *Coffee Break*

Structure Prediction (Chair - Ilan Samish), 10:45 – Noon

10:45 **Ron Unger** (invited presentation): Using Accumulated Dot Matrices to Detect RNA and Protein Structures. (Invited-2)

11:15 **Andreas Heger & Liisa Holm**: Fast Fold Recognition by Consensus Alignment within Transitive Closure. (18)

11:30 **Johannes Soeding** & Andrei N. Lupas: Homology Search By HMM-HMM Comparison Detects More Than Three Times As Many Remote Homologs as PSI-BLAST or HMMER. (41)

11:45 **Jaspreet S. Sodhi**, Kevin Bryson, Liam J. McGuffin, Jonathan J. Ward, Lorenz Wernisch, David T. Jones: Predicting the Location and Identity of Protein Functional Sites, Application to Genomic Fold Recognition. (53)

12:00 *Lunch + Laptop Session*

15:00 **Discussion session** - Challenges in Docking and Structure Prediction. Chair - **David T. Jones**

Structure Analysis (Chair - Kim Henrick), 16:00 - 16:30

16:00 **Eran Eyal**, Marvin Edelman, Vladimir Sobolev: The Influence of Crystal Packing on Protein Structure. (15)

16:15 **Oxana V. Galzitskaya**, Sergiy O. Garbuzynskiy, Bogdan S. Melnik, Michail Y. Lobanov, Alexei V. Finkelstein: Comparison of X-ray and NMR Protein Structures. (22)

16:30 *Coffee Break*

Structure Analysis II (Chair – Vladimir Sobolev), 17:00 - 18:00

17:00 **Maxim Chapovalov & Roland L. Dunbrack**: Using Statistical Analysis of Electron Density to Evaluate Protein Side-Chain Conformations and Rotamer Disorder. (31)

17:15 **Inge Jonassen** & William R. Taylor: A Method for Evaluating Structural Models using Structural Patterns. (47)

17:30 Gregorio Fernandez, Christine Kiel, **Luis Serrano** (invited presentation): In silico Prediction of Protein-Protein Interactions: How and When It Can Work. (Invited-3)

18:00 *Break*

19:00 *Get-together Dinner.*

20:30 **Round Table session** - The Role of Structural Bioinformatics in Computational Biology. **John Moulton** (chair); Panel members - **Geoffrey (Geoff) Barton, Douglas (Doug) Brutlag, Arne Elofsson, John Norvell, Janet Thornton.**

22:00 *Lights out...*

3Dsig Program for Day 2, Friday, July 30th 2004

Non-globular Structures (Chair - **Byungkook (BK) Lee**), 9:00 - 10:15

- 9:00 **Invited Speaker: Gunnar von Heijne**: Genome-wide Membrane Protein Topology Predictions. (Invited-4)
- 9:30 **Matthew L. Baker** & Wah Chiu: Analysis of Intermediate Resolution Structures. (77)
- 9:45 **Marialuisa Pellegrini-Calace**, William R. Taylor, David T. Jones: FILM2: a Novel Method for the Prediction of Transmembrane Helical Bundles in a Membrane-like Environment. (68)
- 10:00 **Rune Linding**, Robert B. Russell, Lars J. Jensen, Francesca Diella, Peer Bork, Toby J. Gibson: Intrinsic Protein Disorder - Function, Disease and Structural Proteomics. (5)
- 10:15 *Coffee Break*

Function Analysis, (Chair - **Haim Wolfson**), 10:45 - Noon

- 10:45 **Invited Speaker: Florencio Pazos & Michael Sternberg**: Automated Prediction of Protein Function from Structure. (Invited-5)
- 11:15 **Richard J. Morris**, Abdullah Kahraman, Rafael Najmanovich, Fabian Glaser, Roman Laskowski, Janet M. Thornton: Protein Active Site Identification and Fast Shape Comparison. (43)
- 11:30 **Antonio Del Sol**, Paul A. O'Meara: Small-World Network Approach to Identify Key Residues in Protein-Protein Interaction. (28)
- 11:45 Arye Shemesh, Gil Amitai, Einat Sitbon, Maxim Shklar, Dvir Netanel, Ilya Venger, **Shmuel Pietrokovski**: Structural Analysis of Residue Interaction Graphs. (45)
- 12:00 *Lunch + Laptop Session*
- 15:00 **Discussion session** - Challenges in Function Prediction. Chair - **Gideon Schreiber**

Molecular Machines and RNA (Chair - **Miriam Eisenstein**), 16:00 - 16:30

- 16:00 **Barry J. Grant**, Leo S. Caves, Rob Cross: Stripping down the Kinesin Molecular Motor: a Combined Informatics and Simulation Approach. (9)
- 16:15 **Joerg Hackermueller**, Nicole-Claudia Meisner, Manfred Auer, Markus Jaritz: mRNA Openers - Computationally Designed Modulators of mRNA Secondary Structure which Manipulate Gene Expression. (74)
- 16:30 *Coffee Break*

Evolution (Chair - **Kevin Karplus**), 17:00 - 18:30

- 17:00 Martin Madera, Julian Gough, Sarah Kummerfeld and **Cyrus Chothia** (invited presentation): The Homology of Genome Sequences: Profile-Profile Sequence Matching and the SUPERFAMILY Database. (Invited-7)
- 17:30 **Sanne Abeln** & Charlotte M. Deane: Investigating Protein Structure Evolution by Fold Usage on Genomes. (24)
- 17:45 Gestel David & **Ian Humphery-Smith**: The Binding-Site Diversity of the Entire Non-denatured, Surface-exposed Human Proteome. (46)
- 18:00 **Invited Speaker: Song Yang**, Russell Doolittle, **Philip E. Bourne**: The Study of Evolution through Protein Domain Structure. (Invited-8)
- 18:30 *Lights out...*

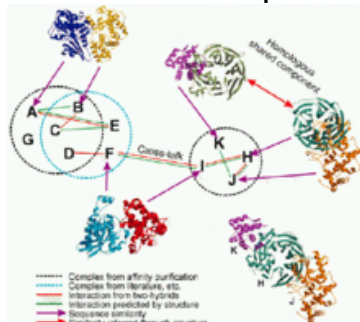
Invited Speaker Abstracts

Number: Invited-1

Authors: Robert B. Russell¹

¹EMBL, russell@embl.de

Title: A Structural Perspective on Protein Interactions & Complexes



Short abstract: I will discuss how protein structure information can be used as a tool to interrogate and predict structures for interactions identified by other methods (e.g. two-hybrids), and present approaches for merging experimental interactions, electron microscopy via structural bioinformatics.

Full abstract: Protein interactions are central to most biological processes, and are currently the subject of great interest. Yet despite the many recently developed methods to identify protein interactions, little attention has been paid to one of the best sources of data: complexes of known three-dimensional (3D) structure. In this talk I will discuss how such complexes can be used to study and predict protein interactions & complexes, and to interrogate interaction networks proposed by methods such as two-hybrid screens or affinity purifications. I will also discuss how EM data can be combined with bioinformatics & modelling to predict structures for complexes or for parts of interaction networks.

Related publications:

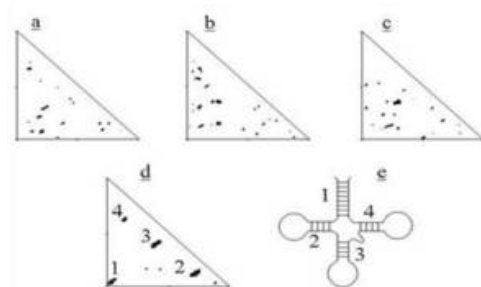
1. P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A.C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, R.B. Russell, Structure-based assembly of protein complexes in yeast, *Science*, 303, 2026-2029, 2004.
2. P. Aloy, H. Ceulemans, A. Stark, R.B. Russell, The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, 332, 989-998, 2003.
3. P. Aloy, F.D. Ciccarelli, C. Leutwein, A.C. Gavin, G. Superti-Furga, P. Bork, B. Boettcher, R.B. Russell A complex prediction: three-dimensional model of the yeast exosome. *EMBO reports*, 3, 628-635, 2002.
4. P. Aloy, R.B. Russell, InterPreTS: Protein interaction prediction through tertiary structure. *Bioinformatics*, 19, 161-162, 2003.
5. P. Aloy, R.B. Russell, The third dimension for protein interactions and complexes. *Trends Biochem. Sci.*, 27, 633-638, 2002.
6. P. Aloy, R.B. Russell Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA*, 99, 5896-5901, 2002.

Number: Invited-2

Authors: Ron Unger¹

¹Bar-Ilan University, ron@biocom1.ls.biu.ac.il

Title: Using Accumulated Dot Matrices to Detect RNA and Protein Structures



Short abstract: We present an approach based on matrix representation of all potential base-pairing of a set of RNA sequences. A proper accumulation of such matrices exposes common structural features, including pseudo-knots. Similar ideas can be extended to analyze aligned protein sequences in order to detect common structural motifs within proteins families.

Full abstract: There is a growing appreciation for the diverse and significant roles RNA molecules play in cellular function. Within this framework we present an approach based on matrix representation of all potential base-pairing of a set of RNA sequences. While a single matrix is too noisy to be used for data extraction, a proper summation of these matrices for related sequences can expose common structural features as demonstrated for tRNA and snoRNA molecules. Furthermore, it is demonstrated, in the case of tmRNA that the method can detect pseudo-knots which are structural motifs that are difficult to detect in other methods. Similar ideas can be extended to analysis of aligned protein sequences in order to detect common structural motifs within proteins families. Within this context we can study long standing questions about the strength of native versus non-native interactions.

Number: Invited-3

Authors: Gregorio Fernandez, Christine Kiel, Luis Serrano¹

¹EMBL, Heidelberg, serrano@EMBL-Heidelberg.DE

Title: In silico Prediction of Protein-Protein Interactions: How and When It Can Work

Full abstract: Determining the interaction repertoire of the proteome of an organism is critical if we want to get an understanding of how complex systems operate. Different experimental techniques have been developed to address this issue and recently procedures to do the same "In Silico" have been published. The "In Silico" procedures are mainly based on the idea that proteins of the same family interact with proteins of another family always on the same way and involving the same surface. This been true then simple contact potentials can be used to discriminate among the multiple possibilities.

In our group we have developed a force field, FOLD-X, that accurately predicts the effect of point mutations in proteins (standard deviation for 1000 mutants of 0.6 Kcal/mol), as well as of complex formation between two proteins. Our strategy to predict genome wide protein interactions is to select domain families for which there is enough structural information of the isolated domains and of the corresponding protein complexes. Then we parse the unknown members of the two domain families over the different isolated template structures and select those that energetically fit on the templates. The selected ones are then used to model the complexes and the interaction energy is determined. We have so far analyzed two types of interactions: SH3 domains with peptide ligands and the Ras-RBD/RA domain interactions. Here we analyze in a critical way the success and failures of the procedure.

Number: Invited-4

Authors: Gunnar von Heijne¹

¹Department of Biochemistry and Biophysics, Stockholm University, gunnar@dbb.su.se

Title: Genome-wide Membrane Protein Topology Predictions

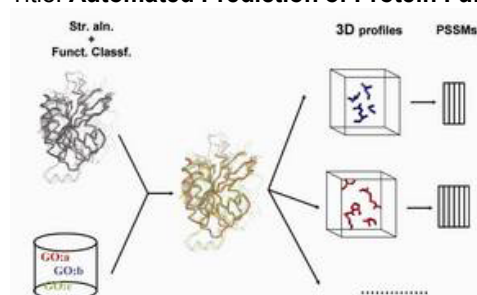
Short abstract: Current estimates are that we can predict the correct topology for 55-60% of all membrane proteins. One way to improve these numbers is to incorporate more experimental data into the predictions. Recent work in our laboratory aiming to produce such data on a genome-wide scale will be presented.

Number: Invited-5

Authors: Florencio Pazos, Michael JE Sternberg¹

¹Imperial College of Science, Technology and Medicine, m.sternberg@imperial.ac.uk

Title: Automated Prediction of Protein Function from Structure



Short abstract: An algorithm is presented that predicts protein function based on structure. The approach uses Gene Ontology Classification (GO) to supervise the identification of functionally-important residues from a structural alignment. In a cross-validated benchmark it is significantly superior to functional transfer from the closest sequence homologues in the twilight-zone for functional prediction (sequence ID=~15%).

Full abstract: With structural genomics projects underway there is a pressing requirement for automated methods to predict function from structure. As a step towards this goal, recently several groups have developed

strategies to identify residues directly associated with function. But actual assignment of function remains problematic. Here we present a new approach that uses the Gene Ontology (GO) classification to supervise the extraction of the sequence signal responsible for protein function from a structure-based sequence alignment. Using GO we can obtain profiles for a range of specificities described in the ontology. Profiles from 121 GO annotations were extracted. The method was tested using cross-validation and compared to the standard method of inheriting the function from the closest homologue. In the region of low sequence similarity (around 15%), our method assigns the correct GO annotation in 90% of the protein tested, around 20% higher than inheritance of function from the closest homologue. The method is also able to identify the specific residues associated by the function of the protein family.

Figure Legend: The Figure provides a schematic diagram of the method used to extract the sequence profile for function prediction.

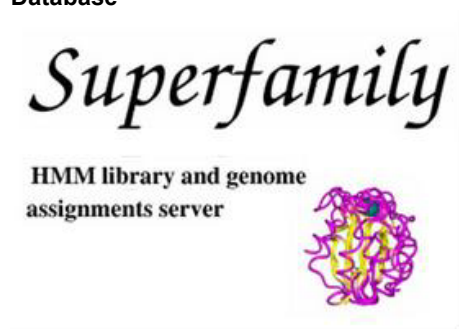
Number: Invited-7

Authors: Martin Madera¹, Julian Gough², Sarah Kummerfeld, Cyrus Chothia³

¹MRC Laboratory of Molecular Biology, Cambridge, UK, mm238@mrc-lmb.cam.ac.uk ³chc1@mrc-lmb.cam.ac.uk

²gough@gsc.riken.go.jp

Title: The Homology of Genome Sequences: Profile-Profile Sequence Matching and the SUPERFAMILY Database



Short abstract: We describe: the features and performance of PProfile Comparer (PRC, <http://supfam.org/PRC>), a new tool for aligning and scoring two profile hidden Markov models; the SUPERFAMILY database (<http://supfam.org>), which lists for all genomes the homologs to domains of known structure, and the distribution of known protein superfamilies in different genomes.

Full abstract: Most protein sequences have evolved through sequence duplication, divergence and recombination: the average size of the known families in the human genome is close to 50. This means that homology determination is central to our understanding of both protein function and the evolutionary processes that have produced the protein repertoires of different organisms.

Profile-profile comparisons are the most powerful method for the detection of sequence homology. PProfile Comparer (PRC) is a new tool for aligning and scoring two profile hidden Markov models. It is probabilistic, makes full use of the transition probabilities encoded by both HMMs, and its alignments and scores are based on a consensus of high-scoring paths through the two models. PRC can read HMMs stored in most common formats. PRC is available for download from <http://supfam.org/PRC> under the GNU General Public Licence. The performance and applications of PRC will be discussed.

The SUPERFAMILY database (<http://supfam.org>), contains the results of matching HMMs for protein domains of known structure against all currently known genome sequences. For the large majority of genomes the matches cover between one third and one half of the residues. Facilities provided by the database will be described. Using data from SUPERFAMILY, we can describe the domain structure and evolutionary relationships of that part of the genome matched by the HMMs. We will describe the number of known protein families in different genomes, the extent to which these families are shared by all organisms, and by the members of different kingdoms.

Number: Invited-8

Authors: Song Yang, Russell Doolittle, Philip E. Bourne¹

¹University of California San Diego, bourne@sdsc.edu

Title: The Study of Evolution through Protein Domain Structure

Short abstract: Consider two principles: 1. Structure can provide information on distant evolutionary relationships. 2. Nature is comprised of a very limited parts list of protein domain folds, hence for a species to lose or gain a fold is a major event. The use of these principles in phylogenetic analysis is discussed.

Full abstract: It is truly remarkable that all of Nature is comprised of a very limited set of parts in the form of protein domain folds - taken here to mean independent and stable folding units. While the exact number remains a point of some conjecture, no one disputes that it is a small number in the 1000's. Since some folds are

promiscuous, taking part in a variety of functions and some are selective, perhaps having a sole function, until we know all biological functions from all organisms we can not be assured of knowing all folds. Nevertheless, at this time we can assign, with reasonable reliability, the number and distribution of domain folds to all fully sequenced genomes (over 150 from the three major kingdoms), achieving 60-80% coverage depending on the complexity of the genome.

This presentation will argue that, based on these assignments, we are at a very interesting time in the study of structural bioinformatics, which classifies domain folds, since a useful contribution can be made to the field of molecular evolution. This statement is based on two underlying principles: 1. structure can provide information on distant evolutionary relationships not seen by current methods of sequence analysis; 2. since Nature is comprised of a very limited parts list of protein domain folds, for a species to lose or gain a fold is a major distinguishing event. Using these two principles we will describe our recent efforts in structure-based phylogenetic analysis, including one interpretation of the tree of life.

Speaker Abstracts

Number: 5

Authors: Rune Linding¹, Robert B. Russell, Lars J. Jensen, Francesca Diella, Peer Bork, Toby J. Gibson

¹EMBL, linding@embl.de

Title: Intrinsic Protein Disorder - Function, Disease and Structural Proteomics.



Short abstract: Disordered regions in proteins often contain short linear peptide motifs (e.g. SH3-ligands and targeting signals) that are important for protein function. Avoiding potentially disordered segments in protein expression constructs can increase expression, foldability and stability of the expressed protein.

Full abstract: A great challenge in the proteomics and structural genomics era is to predict protein structure and function, including identification of those proteins that are partially or wholly unstructured. Disordered regions in proteins often contain short linear peptide motifs (e.g. SH3-ligands and targeting signals) that are important for protein function. We present here DisEMBL, a computational tool for prediction of disordered/unstructured regions within a protein sequence. As no clear definition of disorder exists, we have developed parameters based on several alternative definitions, and introduced a new one based on the concept of "hot loops", i.e. coils with high temperature factors. Avoiding potentially disordered segments in protein expression constructs can increase expression, foldability and stability of the expressed protein. DisEMBL is thus useful for target selection and the design of constructs as needed for many biochemical studies, particularly structural biology and structural genomics projects. The tool is freely available via a web interface (<http://dis.embl.de>) and can be downloaded for use in large scale studies.

The following topics get special attention:

- 1) Protein disorder, concepts and definitions
- 2) Ab-initio prediction of protein disorder
- 3) Functional sites within unstructured regions: linear motifs catalogued by ELM.
- 4) Intrinsically Disordered Proteins such as Tau, Prions, Bcl-2, p53, 4E-BP1 and eIF1A.
- 5) Disorder & Protein Diseases
- 6) HOWTO use disorder predictions to optimise your expression vectors
- 7) Protein beta aggregation and disorder
- 8) The Composite Modular Model of Protein Function
- 9) Demo of WWW servers of ELM, GlobPlot & DisEMBL

REFERENCES

GlobPlot: exploring protein sequences for globularity and disorder. *Nuc. Acids Res.* 2003 - Jul 1;31(13):3701-8.

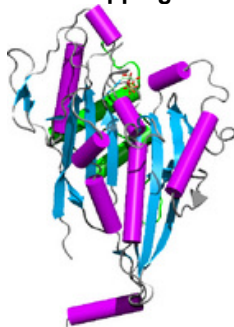
Protein disorder prediction: implications for structural proteomics. *Structure*, 11(11), 2003

ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* 2003 Jul 1;31(13):3625-30.

Number: 9

Authors: Barry J. Grant¹, Leo S. Caves², Rob Cross³

¹University of York & Marie Curie Research Institut, grant@ysbl.york.ac.uk ²University of York, lsdc1@york.ac.uk



Short abstract: Structural informatics and molecular modelling were used to probe sequence-structure-function relationships in the kinesin molecular motor. This study provides information on conformational changes, allosteric modulation, protein-protein interactions and evolutionary relationships. The work illustrates the powerful interplay of informatics and simulation with structural, biochemical and biophysical experiments on this important mechanoenzyme.

Full abstract: Kinesin motors convert the chemical energy from ATP hydrolysis into unidirectional movement along a microtubule track. A principal aim in the study of kinesin, is understanding the mechanism of this energy transduction at the atomic level. Recent advances in experimental methods have afforded exciting new insights into the structure, thermodynamics and kinetics of the underlying processes. However, the data from which we derive knowledge is often incomplete, mismatched in spatial or temporal resolution, and at times, conflicting. Computational modelling and simulation can play an important role in interpreting experimental data and can provide a powerful sounding board for hypotheses concerning aspects of molecular conformation, dynamics and interaction.

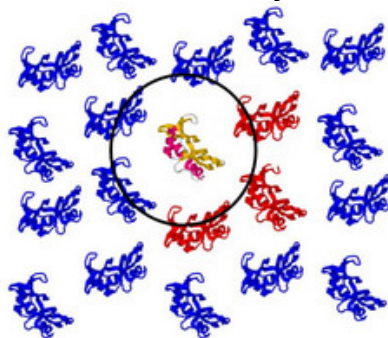
We have performed a number of computational studies on kinesin motors. The available sequence and structural data on kinesin (complemented by data on homologues and analogues) were used to dissect the structural and functional roles of the kinesin architecture. Functionally essential conformational changes were probed by a variety of methods based on empirical potential energy functions, including normal mode analysis (for intrinsic mechanical properties), equilibrium and non-equilibrium molecular dynamics (for the response of systems to changes in the nucleotide state), and Brownian dynamics (for diffusional encounters of the motor with its respective track). We review results from selected studies, illustrating the interplay of simulation with structural, biochemical and biophysical experiments.

Number: 15

Authors: **Eran Eyal¹, Marvin Edelman, Vladimir Sobolev**

¹Weizmann Institute of Science, eran.eyal@weizmann.ac.il

Title: **The Influence of Crystal Packing on Protein structure**



Short abstract: The crystal environment has a small yet statistically significant influence on the structure of a protein. B-factor values of atoms having crystal contacts are smaller. Associated water molecules tend to change position when the environment is changed; associated hetero-groups, less so. See our tool for crystal contact analysis <http://atlantis.weizmann.ac.il/~eyale/cryco>

Full abstract: Abstract: A statistical survey regarding the influence of crystal environment on protein structure has been performed. The environment has been found to have a small but statistically significant influence on backbone as well as side chain conformations. The B factor of atoms having crystal contacts is smaller than of other solvent exposed atoms. Water molecules in the first solvation shell tend to change position when the environment is changed. We did not find an influence of the crystal environment on the position of associated hetero-groups, in agreement with the functionality of many proteins in crystal form. A web based tool for analysis

of crystal contacts is available at <http://atlantis.weizmann.ac.il/~eyale/cryco>.

Introduction: The dense packed environment of globular proteins in crystal structures is dissimilar from the environment in the native state. This raises the question of possible differences between the crystal structure and the structure of the protein in solution. Few large-scale examinations regarding the influence of crystal packing on protein structure have been reported. Most studies deal with side chain conformations (1,2). Usually, side chain modeling programs are evaluated using the coordinates of a single molecule without considering neighbors in the crystal lattice. This might result in under estimating their predictive accuracy, especially for solvent exposed residues. The growth of the PDB database now makes it feasible to extract generalizations regarding structural affects of crystal packing. We have carried out a statistical study on a large number of proteins (more than 500) whose structures were determined at least twice by X-ray crystallography.

Results and discussion: Our main strategy was to compare two structures of the same protein in different crystal forms (i.e., environment). As a control, we used a set of PDB file pairs whose structures have the same crystal environment. While the environment did not normally induce major changes in the fold, we did observe significant differences between pair members in a variety of structural parameters. For atoms is 0.9_ for structures determined in different example, the RMSD of C crystal environments, while 0.4_ in the control set. We investigated if the source of this difference is intra domain structural variability or inter domain relative motion (hinge motion). It was found that in multi-domain proteins in different crystal environments hinge motion frequently occurs.

The water shell around a protein is significantly different between structures with different crystal environments. Water molecules associated with one protein and, additionally, in crystal contact with another, show the largest displacement between structures. Those that are not in crystal contact show smaller displacement but still larger than the water molecules of the control set. This suggests that the entire hydrogen bond network around the protein is significantly different in the two structures. The positions of most other hetero-groups associated with the protein are not effected by the crystal environment, in agreement with the functionality of many proteins in crystal form.

Finally, B factor analysis across our protein set clearly shows that regions of a protein in crystal contact are more rigid, opposite to the conclusion reached for a limited set of five basic pancreatic trypsin inhibitors (BPTI) structures (3). This emphasizes the importance of a large-scale study for this type of question. A tool for analysis of crystal contacts based on contact surface areas is available at <http://atlantis.weizmann.ac.il/cryco>.

References:

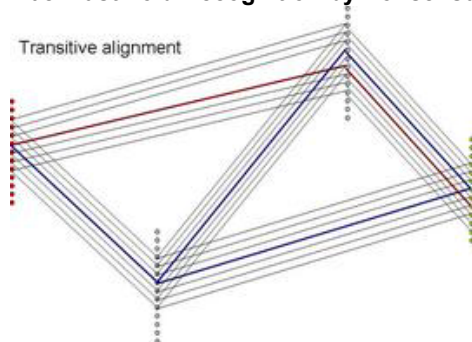
1. Bower MJ, Cohen FE, Dunbrack RL. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol* 1997;267:1268-1282.
2. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol* 2002;320:597-608.
3. Kossiakoff AA, Randal M, Guenot J, Eigenbrot C. Variability of conformations at crystal contacts in BPTI represent true low-energy structures: correspondence among lattice packing and molecular dynamics structures. *Proteins* 1992;14:65-74.

Number: 18

Authors: Andreas Heger¹, Liisa Holm²

University of Helsinki, ¹andreas.heger@helsinki.fi ²liisa.holm@helsinki.fi

Title: Fast Fold Recognition by Consensus Alignment within Transitive Closure



Short abstract: Protein structure is more conserved than sequence. Many distant homologues can only be linked by significant sequence similarity using many intermediate sequences. Pre-processing the information in an all-against-all alignment library enables us to instantaneously generate an optimal transitive alignment between any two proteins, no matter how many intermediates separate them.

Full abstract: Systematic structure comparisons have revealed evolutionarily successful protein super-families which retain a common fold despite sequence divergence beyond recognition. Fortunately, the rapid growth of sequence data means that distant homologues are often linked by more closely spaced intermediate sequences, which are within the detection limit of sequence-based methods. Transitive alignment uses intermediate sequences as stepping stones to infer an alignment between distant homologues. A central problem with transitive alignments is that there are very many choices of intermediates that will lead to mutually inconsistent

alignments between the proteins at the start and end of the chain. We have developed a novel algorithm, MaxFlow [1], which uses a novel consistency score to estimate the relative likelihood of alternative paths of transitive alignment. In contrast to traditional profile models of amino acid preferences, MaxFlow models the probability that two positions are structurally equivalent and retains high information content across large distances in sequence space. Effectively, MaxFlow computes a consensus over transitive alignments based on all-against-all alignments by PSI-Blast. We have shown previously that MaxFlow increases alignment coverage compared to PSI-Blast without compromising reliability. Here, we apply MaxFlow to an input library comprising a representative subset of all known proteins at 40 percent sequence identity level (479,740 sequences). This results in a graph, where most proteins (domains) reside in one huge connected component [2] and means that for any given fold-recognition target sequence, practically all proteins of known structure are potential templates. Fold recognition is achieved by using the MaxFlow alignment score to rank the potential templates (PDB structures). Novel benchmark results on fold recognition specificity and selectivity will be reported.

MaxFlow uses the information from a library of many pairwise alignments to derive an optimal pairwise alignment. Protein sequence evolution is usually modeled using simple statistical models of amino acid preferences at a given position. A sequence is aligned to the model (profile), and the statistical significance of the alignment score is estimated. Thus one can both detect homologues and obtain a sequence-profile alignment with this approach. The profile model can be derived for one sequence (using a generic amino acid substitution matrix) or it can be learned from a set of known homologues (leading to position-specific substitution matrices). The result of a database search is an effective multiple alignment of many sequences against each other, via the profile. Despite being powerful and popular, there is an intrinsic limitation to the profile approach. Profile models are based on log-odds scores for observing a given amino acid type at a given position. From an information theoretical perspective, profile scores “dilute out” as the target distributions of amino acid types observed in structurally equivalent positions broadens at large evolutionary distances and approaches the background distribution. From an evolutionary perspective, the profile model is centred at one point in protein space, and will detect a specific subset of sequence neighbours at some limited radius. However, proteins in different parts of a large super-family evolve under a different set of functional constraints, which affect amino acid preferences and which set of proteins will be detected with statistically significant similarity. Transitive alignment can span very long distances in sequence space by many short steps, exploiting the higher reliability of alignment at shorter evolutionary distances.

Different paths of transitive alignment can lead to mutually inconsistent end points. In the classical multiple sequence alignment problem, one aims to reconcile such inconsistencies using ad hoc objective functions, usually a sum-of-pairs score. This problem is NP-complete and therefore practical algorithms for multiple alignment of large sequence sets have to cut corners or make drastic assumptions. The pairwise alignment problem is solvable exactly given a scoring matrix. MaxFlow’s scoring matrix is derived in a unique way using a novel type of objective function which is based on a path score. The path score is defined between any two proteins in a connected set, even if there is no direct pairwise alignment between the proteins in the underlying pairwise alignment library.

The MaxFlow algorithm is described in detail in ref. [2]. MaxFlow represents the pairwise alignment library as an Alignment Trace Graph, where nodes are residues and edges link equivalent (aligned) residues from two proteins (or domains). Edges are weighted by a consistency score, which reflects the support for the given residue pairing when one considers all possible transitive alignment paths. The path score is defined as the edge weight of the highest weakest link in all paths connecting two residues. To speed up the computation of path scores, MaxFlow clusters residues that have high path scores. These clusters of residues can be thought of as representing columns in a hypothetical multiple alignment. The clustering step is greedy and, for example, does not guarantee that residues are clustered in a consistent sequential order. To recover from such errors, MaxFlow considers the possibility that the optimal pairwise alignment crosses from one hypothetical column (cluster) to another.

MaxFlow evaluates all paths between target and sink residues that reside in the same cluster or neighbouring clusters, where the link is provided by an intermediate residue which is a direct neighbour of either the target or sink residue in the original trace graph. The scoring matrix for pairwise alignment is filled by path scores and the optimal pairwise alignment is determined as usual by dynamic programming.

In conclusion, we report a novel fold recognition tool that uses a single sequence as input. All transitive alignment paths to known structures are evaluated and the highest scoring alignment is returned. Despite computing the transitive closure, the whole procedure is executed in a few minutes on a normal workstation. Beyond fold recognition, the data structures can be used also for many other applications including pattern recognition and data mining.

References

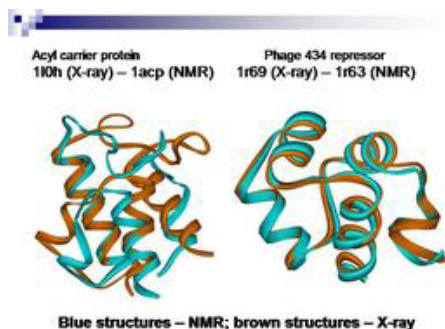
1. Heger A, Lappe M, Holm L (2004) Accurate detection of very sparse sequence motifs. *J Comp Biol*, in press
2. Heger A, Holm L (2003) Exhaustive enumeration of protein domain families. *J Mol Biol* 328, 749-767.

Number: 22

Authors: Oxana V. Galzitskaya¹, Sergiy O. Garbuzynskiy², Bogdan S. Melnik³, Michail Y. Lobanov, Alexei V. Finkelstein

Institute of Protein Research, ¹ogalzit@vega.protres.ru ²S_rgei@mail.ru ³bmelnik@alpha.protres.ru

Title: Comparison of X-ray and NMR Protein Structures



Short abstract: What is the difference between X-ray and NMR-resolved protein structures? A comparison of structures of 61 proteins determined by both NMR and X-ray show statistically reliable differences in the number of contacts per residue and the number of main-chain hydrogen bonds.

Full abstract: Is there a systematic difference between X-ray- and NMR-resolved protein structures? To answer this question, we have compared structures of 61 proteins determined by both NMR and X-ray methods. One of the main found differences between NMR and X-ray structures concerns the number of contacts per residue. We have analyzed the differential picture of the number of contacts per residue at different contact distances (determined as the distances between the closest atoms of residues). We see that: 1) NMR structures have more contacts than X-ray structures at the distances of 2.0 – 3.0 Å and 4.5 – 6.5 Å and less contacts at the distances of 3.0 – 4.5 Å and 6.5 – 8.02 Å; 2) this difference in the number of contacts is greater for internal residues than for external ones; 3) the difference in the number of contacts for internal and external residues depends on the class of proteins: it is larger for beta and alpha+beta proteins than for other classes. Another significant difference between NMR and X-ray structures is the number of main-chain hydrogen bonds, which is smaller for NMR structures than for X-ray structures of the same proteins. On the average, an NMR structure contains only 70% of hydrogen bonds presented by the X-ray structure of the same protein.

Acknowledgements

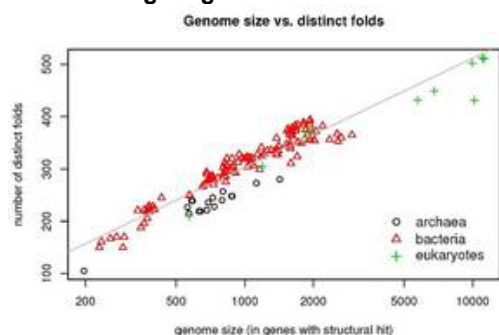
This work was supported by the program MCB RAS, by the Russian Science School program, by the Russian Foundation for Basic Research, by the NWO grant #047.009.021 and by an International Research Scholar's Award to A.V.F. from the Howard Hughes Medical Institute (grant 55000305).

Number: 24

Authors: **Sanne Abeln¹**, **Charlotte M. Deane²**

Department of Statistics, University of Oxford, ¹abeln@stats.ox.ac.uk ²deane@stats.ox.ac.uk

Title: **Investigating Protein Structure Evolution by Fold Usage on Genomes**



Short abstract: Here we use different measures of fold usage on completed genomes in order to explore protein structure evolution. We show that there is not a clear relationship between the number of families per fold, the number of fold copies per genome and the number of genomes a fold occurs on.

Full abstract: Introduction

Here we use the fold usage on completed genomes in order to explore protein structure evolution. We examine the presence or absence of folds on genomes to gain insights into the relationships between folds, the age of different folds and how we have arrived at the set of folds we see today.

Method

Several different measures have been used to describe protein fold usage, such as the number of copies of a fold found on a genome, the number of families per fold and the number of genomes a fold occurs on. We explored all of these categorisations and others by searching for structural domains on 150 completed genome sequences from all three kingdoms of life.

The fold categories used were those from the SCOP database. We tested for their presence using PSI-BLAST on our complete genome set.

We compared the different measures of fold usage and examined the relationships between them. We also considered these properties with different sets of data, e.g. alternating the set of genomes by kingdoms, the set of folds by secondary structure class and substituting folds for super families. We constructed a model and computer simulation that illustrates the distribution of folds or super families over the genomes. The model simulates evolution by generating a species tree from a single genome. Starting with a fixed number of folds on the first genome, folds can either be deleted, copied onto the next genome, or diverge into a new fold on every stage of evolution.

Results & Discussion

In total over 200,000 hits were found on the 150 genomes with an average coverage of 37%. The lowest coverage was 22% for *Borrelia burgdorferi* and the highest 67% for *Buchnera aphidicola*. In our comparisons of these measures we found that bacteria have relatively more distinct folds on their genomes than eukaryotes or archaea (see figure). It was also found that as genome size increased though we did observe an increase in different fold types, this increase varied between the different fold classes (alpha, beta, alpha+beta, alpha/beta).

The expected power law distribution is observed for copies of a fold per genome. We found a similar distribution for the number of families per fold on this set of genomes. However, a more complicated distribution appears for fold occurrence across genomes.

We also show that there is not a clear relationship between the three measures of fold usage. A fold which occurs on many genomes does not necessarily have many copies on each genome. Similarly, folds with many copies do not necessarily have many families or vice versa.

Most observations were similar as we changed from folds to super families, or selected the data from a specific kingdom. However, the distribution of fold occurrence across genomes changes strongly with fold class and kingdom.

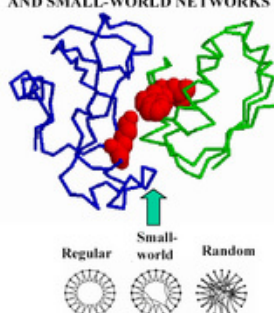
Number: 28

Authors: Antonio Del Sol¹, Paul A. O'Meara²

¹Protein Design Group, National Center for Biotechnology and Research and Development Division, Fujirebio Inc., ao-mesa@fujirebio.co.jp ²Research Division, Fujirebio Inc., pl-omeara@fujirebio.co.jp

Title: Small-World Network Approach to Identify Key Residues in Protein-Protein Interaction.

PROTEIN-PROTEIN INTERACTION
AND SMALL-WORLD NETWORKS



Short abstract: Protein structures are examples of small-world networks, exhibiting a small number of vertices with many connections corresponding to key residues for protein folding. We have successfully used the small-world network approach to predict experimentally verified and new potential candidates for hot spots involved in different biological cases of protein-protein interaction.

Full abstract: Small-world networks have recently become an interesting approach to describe systems such as biological oscillators, genetic control networks, or neural networks. It is characteristic for these networks to be highly clustered, like regular lattices, and to have small characteristic path lengths, like random graphs. Recently, it has been shown that protein structures are examples of small-world networks, exhibiting a relatively small number of vertices with many connections. These vertices have been associated with the central metabolites defining the core of metabolism, key protein domains in proteome evolution, or the key residues for the folding mechanism, among other examples.

Here we have applied the small-world network point of view to study different biological cases of protein-protein interactions (such as hemoglobin and immunoglobulins, proteinase inhibitors, antigens, protease complexes, and hormones) with the aim of identifying the residues, which contribute the most to the binding free energy (hot spots). Although there have been different studies aimed at identifying the key residues which play an important role in protein-protein interactions, the small-world network approach offers the advantage of a global topology point of view of the problem, rather than focusing on localized areas of interest.

We modeled 48 representative protein complexes as un-weighted graphs, and show that they exhibit clustering coefficients and characteristic path lengths corresponding to small-world networks, when compared to random and regular graphs with the same number of nodes and average connectivity.

Next, we looked for the most highly connected residues in the protein complexes, which make the most important contribution in generating the small-world character of these protein complex networks. Our study illustrated that

in 38 of these complexes greater than or equal to 50 percent of these statistically significant highly linked residues occur at the protein-protein interfaces.

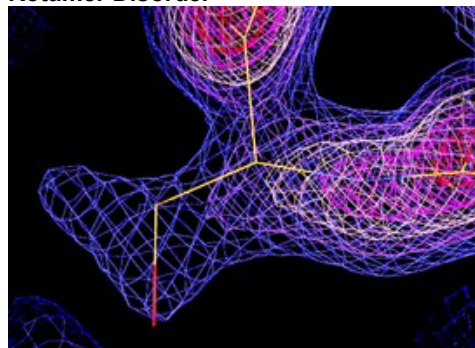
Further analysis on a data set of 19 protein complexes, which contained experimental information on binding free energy hot spots, illustrated that these statistically significant central residues are located in regions characterized by the presence of hot spots. Indeed, 83% of these residues, which are conserved in sequence alignments and non-exposed to the solvent in the protein complex (selected central residues), correspond to or are in direct contact with an experimentally annotated hot spot. The remaining 17% show a general tendency to be close to an annotated hot spot. On the other hand, although there is no available experimental information on their contribution to the binding free energy, detailed analysis of sequence conservation, solvent accessibility in the complex, and correlation between their residue types with the experimental enrichment of hot spot information, shows they are good candidates for being hot spots. Thus, highly connected residues have a clear tendency to be located in regions that include hot spots. In addition, it is shown by a number of examples, that the residue connectivity in the protein complex network is the most appropriate structural characteristic, rather than the residue number of contacts, for identifying binding key residues at the protein-protein interfaces. Interestingly, 11 complexes analyzed are shown to contain at least one selected central residue at its interface, which is also significantly highly connected in the unbound form (overlapping central residues). In all the cases where experimental information on hot spots was available, these overlapping residues correspond to a hot spot. The overlapping residues did not show any substantial difference in terms of their properties respect to the non-overlapping highly central residues in the monomers. Further analysis of these complex interfaces showed that as a general trend in the case of enzyme-inhibitors, the overlapping central residues exhibited a low conformational change respect to the average conformational change of all the interacting residues. Specific topological changes in the monomer structures, which allow their central residues to remain central upon dimerization, are shown through illustrative biological examples. The conformational changes experienced by the interacting residues in the cases of the inhibitor 2bnh involved in the 1dfj endonuclease/inhibitor complex, and the 1f3g phosphocarrier III protein involved in the 1glaGF phosphotransferase complex, are just some of the representative examples studied. This analysis suggests that some information on the important residues for protein-protein association could be extracted from the topology of the monomer structures. In the future, we plan to continue further studies relating to this matter.

Number: 31

Authors: Maxim Chapovalov¹, Roland L. Dunbrack²

Fox Chase Cancer Center, ¹Maxim.Chapovalov@fccc.edu ²RL_Dunbrack@fccc.edu

Title: Using Statistical Analysis of Electron Density to Evaluate Protein Side-Chain Conformations and Rotamer Disorder



Short abstract: Electron density calculations on high-resolution structures demonstrate that non-rotameric side-chain conformations have poor density. Density about chi1 shows that 9% of side-chains exist in more than one chi1 rotamer, with only 2% shown in the PDB entries. These calculations can improve rotamer-libraries and side-chain prediction of single or multiple conformations.

Full abstract: The backbone-dependent rotamer library (<http://dunbrack.fccc.edu/bbdep>) is used in many areas of structure analysis and prediction. A number of side-chain prediction programs, including SCWRL, use the library's backbone-dependent probabilities and dihedral angles. The library is also used in many protein design efforts and in some docking programs that treat side-chain conformational changes in binding. Because the rotamer library is only as good as the data that go into it, we decided to investigate the correlation between side-chain conformations and electron density as calculated from the experimental structure factors in the X-ray experiment.

We obtained a list of 412 high-resolution structures from the PISCES server (<http://dunbrack.fccc.edu/pisces>), from all PDB entries with available structure factors, resolution better than 1.6Å, and R-factor better than 0.20, and no two sequences with sequence identity more than 20%. We used the program CNS (Crystallography and NMR System; Brunger et al. (1998), *Acta. Cryst D* 54:905) to calculate electron densities from structure factors and Cartesian coordinates (for phase determination) downloaded from the Protein Data Bank. We calculated the electron densities at side-chain atomic positions by integrating the electron density in a 1.5 Å-radius sphere around the location of the atom, according to the PDB file, weighted with a decreasing exponential function. These densities were then normalized by dividing by the average density of the backbone, calculated in the same way.

We found that the average density of gamma atoms (atoms CG, SG, OG, OG1, and CG1 in PDB coordinates) varies substantially with the chi1 dihedral angle, and is highest at the staggered (60, 180, and 300 degrees) and lowest at the eclipsed positions (at 0, 120, and 240 degrees) of an sp^3-sp^3 hybridized covalent bond. This result provides some statistical evidence that so-called non-rotameric side chains have very low density and are more likely to be average positions of two interconverting rotameric conformations.

To investigate side-chain disorder further, we rotated a pseudoatom about chi1 and measured the electron density as a function of the dihedral angle. We found four common scenarios. For most side chains, the crystallographic position agrees well with the maximum in the electron density. For a significant fraction of side chains, the crystallographic peak coincided with one electron density peak, usually the maximum, but there was another peak in electron density of nearly the same magnitude. This was especially common for small polar residues, such as serine, with one third of serine residues exhibiting this kind of disorder. In a few cases, a non-rotameric side chain was placed roughly between two distinct rotameric peaks in the electron density. More commonly, a non-rotameric side chain was placed within a large smear of density encompassing two rotameric positions.

Using a conservative criterion for disorder, about 9% of side chains demonstrated multiple occupied rotamers at chi1. This is lower than what is seen in NMR experiments through an analysis of alpha-beta coupling constants (West and Smith (1998), *J. Mol. Biol.* 280:867), which documented a value of 28% in a sample of 10 proteins. However it is much higher than what is documented in the same PDB entries via partial occupancies and multiple coordinates for the XG atom - about 2%. Crystallographic structures are of course less likely to have disorder than NMR structures, due to crystal packing and conditions of temperature, salt, and pH. Crystal packing affects up to 20% of all side chains and up to 60% of surface side chains (Shelenkov and Dunbrack, unpublished). We used electron density estimations at atomic positions from the PDB to determine whether our side-chain prediction program SCWRL3.0 (Canutescu et al. (2003), *Protein Sci.*, 12:2001) predicts side chains with well-determined density better than those with low density. For the top decile of density, SCWRL3.0 predicts 88% of chi1 correctly within 40° of the X-ray structure, while for the bottom decile of density the figure is 66%. This indicates that some of the error in side-chain prediction is likely due to SCWRL predicting one occupied rotamer, while the X-ray coordinates reflect the other occupied rotamer. For whole side chains, RMSD values decrease steadily as electron density increases. This is especially true for the longer side chains. The effect is the smallest in the aromatic side chains as one might expect.

We have developed a system, called Woodchuck, for evaluating side-chain quality with electron density as described above. We have used Woodchuck to identify the top 50% of side chains of each residue type to build a new backbone-dependent rotamer library using a Bayesian statistical method modified from our published work (Dunbrack and Cohen (1997), *Protein Sci.*, 6:1661). This library is built from 1500 structures of 1.8Å or better. As with other criteria for restricting data to the best-determined side chains (Lovell et al. (2000), *Proteins*, 40:389), dihedral angles of the new library have much lower variances and strained rotamers become even rarer than was true of older libraries. Once validated and tested, we plan to use the identification of single and multiple-conformation side chains as a test set for predicting side-chain conformations and rotameric disorder using Boltzmann calculations within SCWRL. Until now, such methods have been tested against multiple structures of the same protein showing different conformations in different PDB entries.

Number: 36

Authors: Maxim Totrov¹

¹Molsoft, max@molsoft.com

Title: Protein-Protein Docking Simulations with Local Backbone Flexibility



Short abstract: Induced fit remains principal obstacle to accurate protein-protein docking. ICM methodology was adapted to efficiently handle full flexibility of segment of polypeptide chain and successfully applied to dock unbound chymotrypsin and eglin with fully flexible 9-residue binding loop. The method may allow accurate docking for systems involving highly flexible loop.

Full abstract: Introduction

While a number of docking algorithms[1] now can easily reconstitute the structure of a protein-protein complex from the subunits taken in their bound conformations, success and accuracy of predictions drops dramatically in the more realistic setup where unbound subunit structures are used[2-4]. Several approaches to the incorporation of side-chain flexibility into the docking algorithms have been reported[5-10]. However, often the conformational

change upon binding is not restricted to side-chains but involves significant backbone movements as well, presenting still bigger challenge. A frequently observed mode of protein-protein association involves binding by one partner of a relatively flexible loop on the surface of the other. Large backbone movements within the binding loop make such systems difficult for protein-protein docking algorithms. For example, experimental structures of bound and free eglin-C and chymotrypsin-alpha are available, and the superposition shows that while bound/unbound heavy atom(backbone) RMSD for chymotrypsin-alpha is only 1.1(0.6)_, for eglin-C it is 2.2(1.4)_, and for its binding loop (residues 42-48) it reaches 4.4(2.9)_. Rigid-body and even flexible side-chain docking fails for this system. Here, locally flexible backbone docking simulation of eglin-C/chymotrypsin-? based on Internal Coordinate Mechanics (ICM)[11-12] is reported.

Method

Generally ICM methodology allows one to keep any part of the system rigid or flexible by selecting the free subset of internal variables, and efficiently includes in the energy calculations only the interactions that depend on these free variables. However, the situation where a fragment of the polypeptide chain needs to be flexible while the adjacent segments are kept rigid poses a problem. Indeed, freeing backbone torsions only within the loop of interest would not automatically restrict flexibility to that region, because the torsions within the loop would also move the C-terminal part of the protein beyond the loop region itself, with respect to the N-terminal part. To circumvent this problem and maintain the efficiency of ICM methodology, the loop of interest was excised from the rest of the protein and represented by a separate internal coordinate tree. Covalent geometry at the junctions was maintained using virtual 'shadow' atoms that duplicated corresponding real atoms in the adjoining part of the molecule – C?alpha atom in the N-terminal part and a C atom in the C-terminal part. Strong harmonic distance restraints between the 'shadow' virtual atoms and corresponding real atoms kept the junctions in a near-ideal configuration. The choice of the junction point between C and C-alpha atoms minimizes the effect of chain disruption on the force-field energy, since the corresponding psi torsion has the smallest rotation barriers among three types of backbone torsions. The rest of the molecule moved as a rigid body, but the side-chain torsions of the residues in the vicinity of the loop and the C-terminal glycine residue were also left flexible.

ICM biased probability Monte-Carlo (BPMC)[13] global minimization procedure was applied. Free side-chain torsions and loop backbone phi and psi angles were subject to BPMC random moves, each followed by up to 1500 steps of local gradient minimization including the variables controlling the overall position and orientation of the ligand. Simulation was terminated after 10 million energy evaluations, or approximately 17000 random moves.

Initial position of the ligand (eglin-C) was taken from the rigid-body grid docking simulation previously reported[4]. Conformation which had the binding loop of ligand oriented towards the binding site of receptor was selected as a starting point for flexible docking simulation. This conformation had RMSD of 5.3_ for the heavy atoms of the loop residues (see Figure). While the mutual orientation of the ligand and receptor resembled that of the native complex, the interface was very loose and residue-residue contacts were mostly wrong due to a completely different conformation of the loop in the NMR structure of unbound eglin-C (PDB code 1egl) used as a starting point.

The receptor (chymotrypsin-alpha) molecule was taken from an unbound X-ray structure (PDB code 5cha). In the simulations it was represented by grid potentials as described in ref. 4. While flexibility of the receptor was not treated explicitly, the use of truncated van der Waals potentials (maximal repulsion 1.5 kcal/M) implicitly allowed for some 'softness'.

Results and discussion

Ten simulations were run in parallel to ensure convergence. 3 runs achieved very similar lowest energy conformations (see Figure).

The best energy conformation had all side-chains of the binding loop correctly placed into the corresponding receptor pockets, and RMSD for the 7 loop contact residues with respect to the experimental complex structure (PDB code 1acb) was only 1.9_1.3)_ for the backbone).

Thus, significant conformational changes can be successfully predicted by locally flexible backbone docking simulations in internal coordinates. In the present study, it was assumed that the flexible binding loop is known. While this requirement is clearly a limitation, such information may often be available from the non-structural experimental information (mutations), structural data for free subunits (b-factors, disordered regions), and from computational evaluation of local protein flexibility. The results demonstrate feasibility of inclusion of the backbone flexibility into protein-protein binding simulations. Further test s on systems including protein/ligand complexes are underway.

References:

1. Halperin I, Ma B, Wolfson H, Nussinov R. *Proteins*. 2002 Jun 1;47(4):409-43.
2. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. *Proteins*. 2003 Jul 1;52(1):2-9.
3. Mendez R, Leplae R, De Maria L, Wodak SJ. *Proteins*. 2003 Jul 1;52(1):51-67.
4. Fernandez-Recio J, Totrov M, Abagyan R. *Protein Sci*. 2002 Feb;11(2):280-91.
5. Totrov M, Abagyan R. *Nat Struct Biol*. 1994 Apr;1(4):259-63.
6. Jackson RM, Gabb HA, Sternberg MJ. *J Mol Biol*. 1998 Feb 13;276(1):265-85.
7. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. *J Mol Biol*. 2003 Aug 1;331(1):281-99.
8. Zacharias M. *Protein Sci*. 2003 Jun;12(6):1271-82.
9. Lorber DM, Udo MK, Shoichet BK. *Protein Sci*. 2002 Jun;11(6):1393-408.

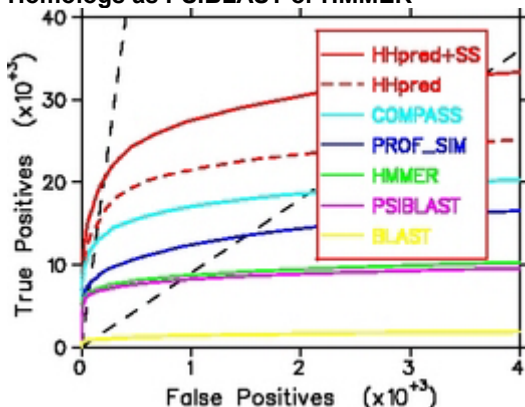
10. Fernandez-Recio J, Totrov M, Abagyan R. *Proteins*. 2003 Jul 1;52(1):113-7.
11. Abagyan RA, Totrov MM, Kuznetsov DA. *J Comp Chem* 1994, 15, 488-506.
12. MolSoft ICM 3.0 program manual; MolSoft LLC: San Diego, 2003. On-line:
<http://www.molsoft.com/services/help/frames.htm>
13. Abagyan R, Totrov M. *J Mol Biol*. 1994 Jan 21;235(3):983-1002.

Number: 41

Authors: **Johannes Soeding¹**, **Andrei N. Lupas²**

Max-Planck-Institute for Developmental Biology, Dept Protein Evolution, ¹johannes.soeding@tuebingen.mpg.de
²andrei.lupas@tuebingen.mpg.de

Title: **Homology Search By HMM-HMM Comparison Detects More Than Three Times As Many Remote Homologs as PSIBLAST or HMMER**



Short abstract: We developed a homology search method based on pairwise comparison of profile HMMs. In an all-against-all benchmark (3691 SCOP 1.63 domains, < 20% sequence identity) the method detects 3-5 times more homologs than PSIBLAST or HMMER and between 25% and 100% more than profile-profile comparison tools COMPASS and PROF_SIM.

Full abstract: A very successful strategy for protein structure prediction relies on identifying homologous sequences with known structure to be used as template. Sensitivity in homology detection is crucial since many proteins have only remote relatives in the structure database. Using as much information about the query and target proteins as possible should allow a better distinction of true from false positives and result in a better alignment quality. This is the reason why sequence-sequence comparison is inferior to profile-sequence comparison. Similarly, methods based on profile-sequence comparison are clearly outperformed by profile-profile comparison methods. Profile Hidden Markov Models (HMMs) are similar to simple sequence profiles, but in addition to the amino acid frequencies in the columns of a multiple sequence alignment they contain information about the frequency of inserts and deletions at each column. Using profile HMMs in place of simple sequence profiles should further improve sensitivity.

We have developed a method for sequence homology search and alignment which is based on the pairwise comparison of profile HMMs. The alignment algorithm maximizes a weighted form of coemission probability, the probability that the two HMMs will emit the same sequence of residues. Amino acids are weighted according to their abundance, rare coemitted amino acids contributing more to the alignment score. This weighting is analogous to the use of a null model with amino acid background probabilities in HMM-sequence comparison. Secondary structure can optionally be included in the HMM-HMM comparison. We score pairs of aligned secondary structure states in a way analogous to the classical amino acids substitution matrices. We use ten different substitution matrices that we derived from a statistical analysis of the structure database, one for each confidence value given by PSIPRED. By making use of confidence values and using a strictly statistical approach we are able to improve the sensitivity gain considerably over previous methods.

Our tool HHsearch is benchmarked together with BLAST, PSI-BLAST, HMMER, and two profile-profile comparison tools, PROF_SIM and COMPASS, in an all-against all comparison of a large database of 3691 structural domains from SCOP 1.63 with less than 20% sequence identity. Depending on the definition of true and false positives, HHsearch is able to detect between three and five times more homologs than PSIBLAST or HMMER and between 25% and 100% more than COMPASS or PROF_SIM (at a fixed error rate of 10%). By including secondary structure predicted by PSIPRED a further improvement between 20% and 40% could be achieved. We show that the position-specific gap information and secondary structure also improves the quality of alignments. Thanks to a fast column score and an efficient algorithm, HHsearch runs only 2.3 times slower than HMMER, 10 times faster than PROF_SIM, and 17 times faster than COMPASS, scanning a profile HMM with 200 columns against the 3691 SCOP domains in 33 s on a 3 GHz AMD64 processor.

The large number of compared sequences in our benchmark allows us to quantitatively test the accuracy of reported E-values. Remarkably, E-values below 1E-4 are wrong for all tested tools except BLAST and HMMER. Specifically, when using PSI-BLAST with standard parameters (up to eight iterations at E-value threshold of 1E-4) ~100 times too many sequences were observed with a reported E-value of 1E-4 or better. In order to give the user a reliable estimate of significance, HHsearch calculates the probability of a hit being a true positive, based

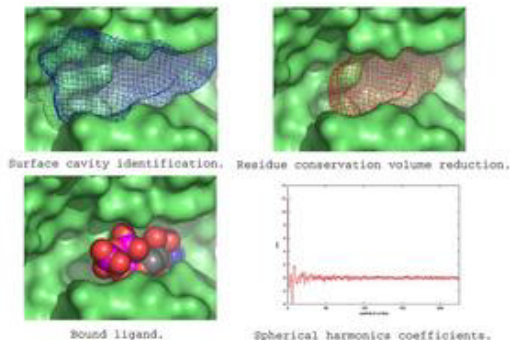
on the real-world score distribution for negative and positive domain pairs. HHsearch and the database of alignments and profile HMMs used in the benchmark can be downloaded from our website at <http://protevo.eb.tuebingen.mpg.de/download/>. HHPred is a structure prediction server based on HHsearch that is available on the web via <http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred>

Number: 43

Authors: Richard J. Morris¹, Abdullah Kahraman², Rafael Najmanovich, Fabian Glaser, Roman Laskowski, Janet M. Thornton

EBI, ¹rjmorris@ebi.ac.uk ²abdullah@ebi.ac.uk

Title: Protein Active Site Identification and Fast Shape Comparison



Short abstract: Methods are presented to aid protein function prediction from structure. We focus on algorithms to locate and to describe a protein's active site with parameters from a multipole expansion. An active site similarity metric is presented that allows for rapid comparisons against a large signature database.

Full abstract: We present a method capable of capturing the overall shape of a protein's active site and also the mapping of physical-chemical properties onto the protein surface. The approach uses the methodology of SURFNET combined with a phylogenetic-based conservation algorithm to predict the location and shape of potential ligand binding sites, we then employ a spectral decomposition technique (spherical harmonics expansion) to describe the shape. The same technique can be applied to model other properties such as electrostatic charge. These expansion parameters are then used to build a multi-dimensional feature vector. Close points in this feature space also correspond to similar active sites. This allow extremely fast comparisons to be carried out based on feature vector products (or differences).

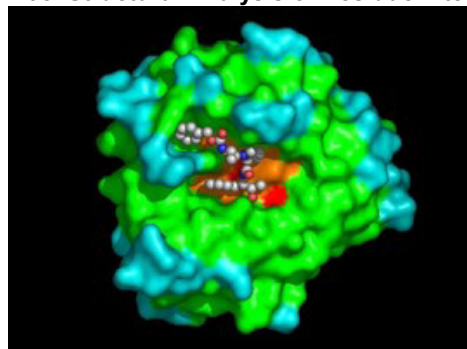
Figure Legend: Potential binding sites are identified with Surfnet and then reduced to the vicinity of conserved residues. The resulting shape is modelled with spherical harmonics and compared to other proteins and ligands.

Number: 45

Authors: Arye Shemesh¹, Gil Amitai, Einat Sitbon, Maxim Shklar, Dvir Netanel, Ilya Venger, Shmuel Pietrokovski

¹Weizmann Institute of Science, arye.shemesh@weizmann.ac.il

Title: Structural Analysis of Residue Interaction Graphs



Short abstract: We present a graph theory approach for studying protein structures. Structures were transformed to residue interaction graphs, where nodes are residues and edges connect interacting residues. Centrality measures of nodes identified various functional sites. Our method analyzes single structures, is independent of conservation, and does not rely on prior training.

Full abstract: Prediction of protein functional sites from their structure alone is an important and difficult task. New structures are solved at an increasing rate, with many having only poor functional documentation. Structure and sequence similarity facilitates the identification of functional sites. However, many proteins include unique structure folds (unifolds), or have no known sequence homologs (ORFans).

We present a graph theory approach for analyzing protein structures. Protein structures are transformed to residue interaction graphs (RIGs), where nodes are amino acid residues and edges connect interacting residues. Our results show that residues with high centrality values are typically found in functional sites (e.g., catalytic or ligand binding sites). These centrality values are independent of other residue structural measures, such as 'clefiness' or B-factor.

The centrality measure we used is determined by the relation of each node (residue) to the whole network (protein structure). This is a global measure, as opposed to local ones such as node connectivity. As expected, residues at the protein core have high centrality values relative to those at the protein surface. However, functional sites, residing on the protein surface, have even higher centrality values. Residues with high centrality values are also evolutionarily conserved, and are more susceptible to deleterious mutations. Both these relations were found to be statistically significant.

A large scale prediction of active sites in 178 representative protein structures, using only residue centrality values, gave on average 54% sensitivity and 5% selectivity. Adding the residues' relative solvent accessibility gave on average 46% sensitivity and 10% selectivity. These results are slightly better than other existing prediction methods, which use more structural properties.

Our method of functional sites identification has several unique advantages over other approaches. RIG analysis can be done on a single structure, is independent from sequence information, and does not rely on prior training. Moreover, since it uses a new and independent property of protein structure, it can be combined with other sequence and structure measures (sequence conservation, residue charge, amino acid catalytic propensity, structure topology etc.) to give an even stronger integrated predictive method.

Graph analysis for prediction of protein functional sites is a new concept. As such, it gives us new types of measures to analyze protein structures. Specifically, node centrality was shown to correlate with functionality of protein residues. Studying the theoretical basis for the success of RIG analysis presents us with a new view of the complex relations between protein structure and function.

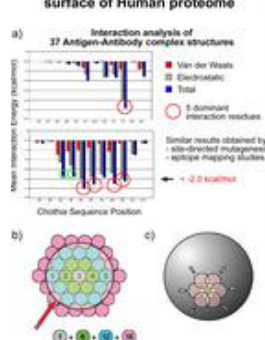
Number: 46

Authors: Gestel David¹, Ian Humphery-Smith²

Dept. of Pharmaceutical Proteomics, Universiteit Utrecht, ¹david.gestel@pharm.uu.nl ²ianhs@hotmail.com

Title: The Binding-Site Diversity of the Entire Non-denatured, Surface-exposed Human Proteome

Mean interaction patch walked across surface of Human proteome



Short abstract: Thirty-seven antigen / antibody crystal structures were employed to establish a mean interaction patch radius. A minimalist model corrected with data derived from 680 SCOP domain structures was then employed in association with every Open Reading Frame in the human genome to provide a reliable estimate of potential binding-site diversity.

Full abstract: Antibodies are highly non-globular proteins. Their epitope-binding domain is therefore an example of a smaller than normal interaction patch. Nonetheless, 37 antigen/antibody crystal structures were employed to establish a reliable estimate of a mean interaction patch radius as determined by Van de Waals and electrostatic forces. A minimalist model for surface-exposed residues was then corrected with real data derived from 680 SCOP domain structures. Each interaction patch corresponded to 2,408 potential and distinct five-mer interactions once disc overlap was subtracted. When the interaction patch was walked across the surface of all proteins encoded by every Open Reading Frame in the human genome it resulted in approximately 20 billion potential interactions. The latter figure took into account reliable estimates of the number of protein isoforms engendered by post-translational modifications and splice variants. As in computer science, biological systems must set about reducing this potential dimensionality so as to undertake a meaningful intracellular work. Factors involved in reducing dimensionality include the concentration of both targets and ligand, on-and off-rates, spatial and temporal segregation, molecular mass, the number of binders in a biological complex and combinatorial effects resulting in physiological thresholds as an endpoint of one or more reaction pathways.

A number of physiological consequences could be drawn from these calculations of relevance to protein-protein interactions in general and the establishment of self-tolerance in mammals. In silico predictions clearly demonstrated a fantastically-high potential for linear MHC-peptides to cross-react with conformational 'self', which if not addressed during lymphocyte ontogeny has important consequences for 'non-self' recognition and autoimmunity mediated by T-Cell receptors; B-Cell receptors and antibodies. The minimal information threshold required for self / non-self recognition by MHC-peptides was accurately determined, as was the obligatory nature of polyclonality for both cellular and humoral immunity. The latter is likely to be governed by similar mathematical

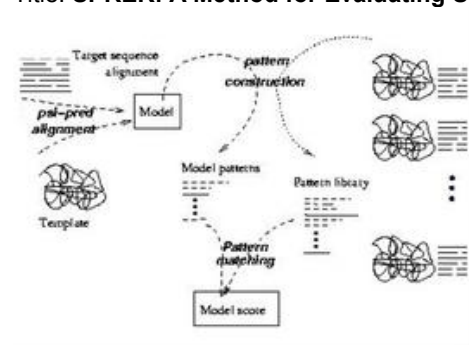
constraints as bioinformatic tools employed in proteomics, namely 'combinatory sieving'. At the molecular level, it is portions of populations of molecules and physiological thresholds attained by one or more pathway end-products that will result in 'specificity emergent'.

Number: 47

Authors: Inge Jonassen¹, William R. Taylor²

¹University of Bergen, inge@ii.uib.no ²National Institute of Medical Research, London, wtaylor@nimr.mrc.ac.uk

Title: SPREK: A Method for Evaluating Structural Models using Structural Patterns



Short abstract: We propose a novel method for the evaluation of structural models. The captures multiple interactions and rewards models having interaction patterns frequently found in true structures. The method utilises string representations of local neighbourhoods. A comparative evaluation of SPREK demonstrates that it is competitive with state of the art methods.

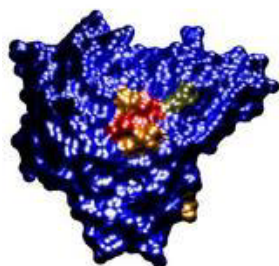
Full abstract: A method (SPREK) was developed to evaluate the register of a sequence on a structure based on the matching of structural patterns against a library derived from the protein structure databank. The scores obtained were normalised against random background distributions derived from sequence shuffling and permutation methods. Random structures were also used to evaluate the power of the method. These were generated by a simple random-walk and a more sophisticated structure prediction method that produced protein-like folds. For comparison with other methods, the performance of the method was assessed using collections of models including decoys and models from the CASP-5 exercise. The performance of SPREK on the decoy models was equivalent to (and sometimes better than) those obtained with more complex approaches. An exception was the two smallest proteins where SPREK did not perform well due to a lack of patterns. Using the best parameter combinations from trials on decoy models, the CASP models of intermediate difficulty were evaluated by SPREK and the quality of the top scoring model was evaluated by its CASP ranking. Of the 14 targets in this class, half lie in the top 10 percent (out of around 140 models for each target). The two worst rankings resulted from the selection by our method of a well packed model that was based on the wrong fold. Of the other poor rankings, one was the smallest protein and the others were the four largest (all over 250 residues).

Number: 53

Authors: Jaspreet S. Sodhi¹, Kevin Bryson², Liam J. McGuffin, Jonathan J. Ward, Lornenz Wernisch, David T. Jones

University College London, ¹j.sodhi@cs.ucl.ac.uk ²K.Bryson@cs.ucl.ac.uk

Title: Predicting the Location and Identity of Protein Functional Sites, Application to Genomic Fold Recognition



Short abstract: We present a fully automatic site detection method, FuncSite, that uses neural network classifiers to predict the location and type of functionally important sites in protein structures. We have shown the method to be robust enough for site detection in low resolution structural models applicable to genomic fold recognition predictions.

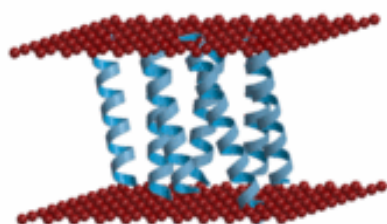
Full abstract: World-wide structural genomics initiatives are rapidly accumulating structures for which limited functional information is available. Additionally, state-of-the art structural prediction programs are now capable of generating at least low resolution structural models of target proteins. Accurate detection and classification of functional sites within both solved and modelled protein structures therefore represents an important challenge. We present a fully automatic site detection method, FuncSite, that uses neural network classifiers to predict the location and type of functionally important sites in protein structures. The method is designed primarily to require only relative residue position without the need for specific side-chain atoms to be present. The functional site encoding represents conservation using PSI-BLAST PSSMs of site residues as well as solvent accessibility and secondary structure assignments. We have rigorously benchmarked FuncSite on a set of metal binding sites spanning numerous SCOP super-families. The method has also been extended to the prediction of protein-DNA interface regions, adenylate classification and the identification of enzyme active sites. In order to highlight effective site detection in low resolution structural models FuncSite was used to screen model proteins generated using mGenTHREADER on a set of newly released structures. We found effective metal site detection even for moderate quality protein models illustrating the robustness of the method. Previously the Genomic Threading Database (GTD) was developed to obtain structural predictions for a wide range of organisms (McGuffin et. al.). This repository was screened in order to identify likely functional sites with the aim of extending current functional annotations. We have also investigated the use of site detection to improve fold recognition predictions. The method applies site score in order to distinguish accurate protein models from a set of predictions.

Number: 68

Authors: Marialuisa Pellegrini-Calace¹, William R. Taylor, David T. Jones

¹EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kin, marial@ebi.ac.uk

Title: FILM2: a Novel Method for the Prediction of Transmembrane Helical Bundles in a Membrane-like Environment



Short abstract: A novel method for the prediction of helical transmembrane protein structures, FILM2, is presented. The method combines the helix variphobicity, which predicts the exposure of each transmembrane helix to lipids, with a previously developed knowledge-based membrane potential and can predict with reasonable accuracy both helix topologies and the protein folds.

Full abstract: Membrane proteins make up a wide and important class of biological macromolecules, representing about 30% of the expression products of the genes predicted from genome sequencing, and have been shown to play a key role in the regulation of selective transports across cell membranes [1]. Despite that and the continuous improvement of biophysical method power and speed, still less than 2% (530 vs 25,551) of the 3-D structures deposited at the Protein Brookhaven Data Bank (PDB) are membrane-associated proteins [2] and only few attempts have been made to model membrane proteins at 3-D level using primary sequence information alone.

We recently developed FILM, a knowledge-based method for the prediction of small membrane protein structures based on the addition of a membrane potential to the energy terms of a previously developed ab initio technique for the prediction of globular protein folds [3]. The membrane potential was derived by the statistical analysis of 640 transmembrane helices with experimentally defined topology from the SWISS-PROT database, while the fold generation itself was based on the assembly of supersecondary structural fragments taken from a library of highly resolved protein structures using a standard simulated annealing algorithm.

Herein FILM2, a new version of the method suitable for larger helical transmembrane proteins, is presented. The main innovation consists in the introduction of a further membrane-related parameter, the helix variphobicity [4]. Just as exposure to solvent in globular proteins can be characterised from sequence variation, the prediction of the variphobicity of transmembrane helices is based on the substitution in a multiple sequence alignment of variable hydrophilic positions (indicators of exposure to solvent) with variable hydrophobic (variphobic) positions and indicates the degree of exposure to lipid. Moreover the previously used fold generation approach was replaced by a combinatoric enumeration of windings over a predefined architecture [4].

FILM2 is fully automated from the initial sequence data bank searches to the final construction of 3-D models and includes several steps:

- a) generation of a multiple alignment for the target protein by a BLAST search [5];
- b) prediction of the number and the orientation of transmembrane helices in the target protein by MEMSAT [6-7];

- c) prediction of the size of the variphobic (i.e. evolutionarily variable and hydrophobic) faces of transmembrane helices;
- d) generation of numerous folds by a combinatoric enumeration of windings over a predefined architecture and selection of a number of preferred alternatives according to the predicted variphobicity values;
- e) evaluation of the selected folds by a knowledge-based potential composed by pairwise, solvation, hydrogen-bonding and membrane potential energy terms.

The method was successfully applied to few protein family sequence alignments strongly differing in extent and degree of internal similarity among the sequences and results highlighted that FILM2 is able to predict with reasonable accuracy both the helix topology and the conformations of these proteins.

References

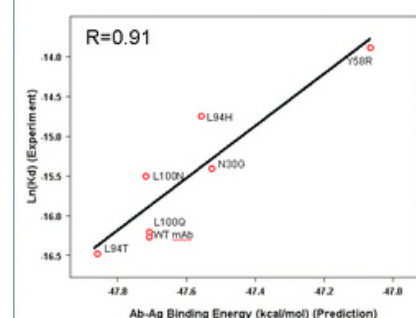
- [1]. Wallin E, von Heijne, G. "Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms". *Protein Sci* 1998, 7, 1029-1038.
- [2]. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. "The Protein Data Bank." *Nucl Ac Res* 2000, 28, 235-242.
- [3]. Pellegrini-Calace M., Carotti A., Jones D.T. "Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures." *Prot. Struct. Funct. Gen.* 2003, 50, 537-545.
- [4]. Taylor W.R., Jones D.T., Green N.M. "A method for alpha-helical integral membrane protein fold prediction." *Prot. Struct. Funct. Gen.* 1994, 18, 281-294.
- [5]. Altschul S.F., Madden T.L., Schnffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.*, 1997, 25, 3389-3402.
- [6]. Jones D.T. "Do transmembrane protein superfolds exist?" *FEBS letters*, 1998, 423, 281-285.
- [7]. Jones D.T., Taylor W.R., Thornton J.M. "A Model Recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology." *Biochem.*, 1994, 33, 3038-3049.

Number: 72

Authors: Qiaojuan Jane Su¹, Orit Foord², Scott Klakamp, Holly Tao, Larry L. Green

¹Abgenix, jane.su@abgenix.com ²orit.foord@abgenix.com

Title: Modeling and Simulation of Antibody-Antigen Interaction: An Integrated Approach



Short abstract: We have built an antibody-epitope docking model that integrates epitope mapping and affinity data. Using this model, we show that structure-based antibody-antigen binding energetics simulations can correlate well with experimental affinity ranking. We propose that a proven docking model might further be applied to perform *in silico* design to optimize antibody characteristics.

Full abstract: The immune system is able to generate a large and diverse repertoire of antibodies recognizing an expanse of epitopes. Antibodies can demonstrate extraordinary affinity and specificity toward their antigens, and thus are invaluable as molecular probes for research and more recently as diagnostic and therapeutic agents. Modeling of the antibody combining site provides insights into the molecular basis of antibody-antigen interaction, and forms the working hypothesis for rational design and engineering of an optimized antibody. The approach and principles discovered can be extended to more general ligand-receptor or protein-protein interaction studies. EGFRvIII is a tumor-specific variant of the human epidermal growth factor receptor (EGFR). We have generated a panel of fully human EGFRvIII monoclonal antibodies using Abgenix's proprietary XenoMouse[®] technology. High-affinity human monoclonal antibodies specific to EGFRvIII are candidates for conjugation to cytotoxic drugs for use as therapeutics against solid tumors. We have built a three-dimensional structural model of the variable region of a fully-human EGFRvIII antibody through a homology modeling approach using the InsightII molecular modeling package (Accelrys). The antibody model illustrates a long and narrow binding groove formed by the complementarity-determining regions (CDR).

A series of experiments were performed to characterize the epitope on EGFRvIII. The fine-resolution epitope mapping studies revealed that a continuous seven-residue peptide contains key binding residues on the antigen. Therefore, this seven-residue peptide was docked onto the antigen-binding site on the antibody structure model. The docking models were built with the Docking module in InsightII.

First, a structural model for the seven-residue peptide was built in an extended conformation and energy minimized with Discover_3 module. Next, the peptide structure was manually placed into the combining site to form an initial assembly. A Monte Carlo search in the translational and rotational space was then automatically performed in the Docking module. The plausible configurations found by Monte Carlo search were subjected to simulated annealing and energy minimization to reach the final models. A total of 63 docking models were

obtained. For each docking model, the interaction energy between the antibody and the individual residue in the peptide was calculated. The profiles of individual residue contribution in epitope-antibody binding were inspected to select the docking models that are consistent with the epitope mapping data. 19 out of 63 docking models have energy profiles consistent with the experimentally identified key binding residues.

The antibody residues at the binding interface comprise the paratope. Mutation of the paratope residues has various impacts on epitope-paratope interaction. To simulate the sequence-affinity relationship, we carried out *in silico* mutations for all the paratope residues and calculated the antibody-antigen binding energy for each mutant. Therefore, we obtained predicted affinity ranking tables derived from each of the 19 docking models.

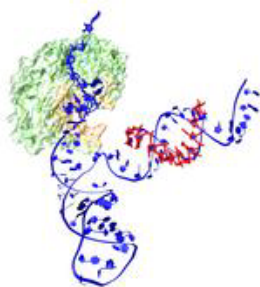
Experimentally, we selected a number of mutations to perform site-directed mutagenesis and determined their affinities to the peptide with Biacore. For each model, a linear regression analysis was done for the predicted antibody-antigen binding energies against the logarithm of experimentally determined dissociation constants. A wide range of correlation coefficients was obtained from the 19 docking models, with the best having a correlation coefficient of 0.91. The docking model giving this excellent correlation is thus our best docking model. Together with the validated affinity-ranking model, this docking model might further be applied to optimize antibody characteristics. Therefore, with a reliable antibody-antigen complex structure model, *in silico* antibody selection is promising.

Number: 74

Authors: Joerg Hackermueller¹, Nicole-Claudia Meisner², Manfred Auer, Markus Jaritz

Novartis Institute for Biomedical Research Vienna, ¹joerg.hackermueller@pharma.novartis.com ²nicole-claudia.meisner@pharma.novartis.com

Title: mRNA Openers – Computationally Designed Modulators of mRNA Secondary Structure which Manipulate Gene Expression



Short abstract: Recognition of RNA secondary structure motifs is of central importance for regulatory RNA protein interactions. The presented method allows to explain measured RNA protein affinities by quantitatively describing RNA secondary structure motif availability. Computationally designed secondary structure modulators boost or inhibit regulatory RNA protein interactions as predicted.

Full abstract: RNA secondary structure is of central importance in the regulation of many cellular processes such as pre-mRNA processing, metabolic pathways or other control mechanisms depending on RNA protein interactions. In particular, we have recently shown that it is the secondary structure which controls mRNA stability via interaction with the RNA binding protein HuR. This tightly controlled process determines the expression of approximately 3,000 genes, including disease relevant proteins and pharmaceutical targets in cancer, inflammatory, viral, allergic, vascular and infectious diseases. Hence, modulating such regulatory RNA-protein interactions through mRNA secondary structure manipulation may serve as a novel strategy for therapeutic intervention.

We have developed a computational method to model the RNA secondary structure influence on the recognition by a particular protein. We assumed that the protein would bind only to RNA molecules which form a particular secondary structure motif and deduced that in this case RNA protein affinities depend on the availability (i.e. the concentration) of the motif. We therefore quantify the availability of a secondary structure motif as the thermodynamic probability of secondary structures in the ensemble which comply with a particular secondary structure motif. This quantitative measure of the motif availability correlates well with experimentally measured affinities of a regulatory protein to its target mRNAs. Using this method we were able to dissect the molecular mechanism of HuR target mRNA recognition and showed that HuR requires the presentation of a degenerate 9mer sequence motif in single stranded conformation - analogous to a regulatory on/off switch which controls HuR access to mRNAs *in cis*.

Based on this approach we computationally designed short oligonucleotides for a directed manipulation of mRNA secondary structure and tested their application for specifically modulating gene expression. We predicted the effect of the hybridization of these oligonucleotides with a target mRNA on the availability of a motif crucial for a particular protein interaction. Such modulators not only allow to e.g. unfold a regulatory hairpin but to disrupt or favor any functional RNA structure element. Such an artificially induced conformational reorganization - occurring at regions proximal or distant to the hybridization site within the mRNA sequence - allows to hide or present the recognition site of a regulatory factor. It can thereby be used to manipulate the associated regulatory process, e.g. controlling the expression level of a particular gene. We calculated the impact of hybridization on the availability of a motif for all oligonucleotides of fixed, arbitrarily selected length which are exactly reverse

complementary to a target RNA. Oligonucleotides which locally maximize or minimize the probability of the secondary structure element upon hybridization were selected for experimental validation. In the case of HuR associated mRNA stabilization we were able to design modulators which increased in vitro and endogenous HuR-mRNA complex formation and demonstrated the associated specific mRNA stabilization independently for two different cytokine mRNAs.

The methods therefore provide a means for (i) dissecting molecular mechanisms of secondary structure dependent RNA – protein interactions, (ii) the development of target specific assays and high throughput screens for compounds which inhibit or strengthen the modulator induced structure change on a particular target mRNA, (iii) for silencing or boosting gene expression in cell biological or in vivo assays as an alternative to siRNA and (iv) for mechanistic studies of the biological role of one particular protein binding site by specifically “opening” or “closing” an individual among multiple binding sites.

Number: 76

Authors: Yuval Inbar¹, Haim J. Wolfson², Ruth Nussinov

Tel Aviv University, ¹inbaryuv@tau.ac.il ²wolfson@tau.ac.il

Title: Prediction of Multi-Molecular Assemblies by Multiple Combinatorial Docking.

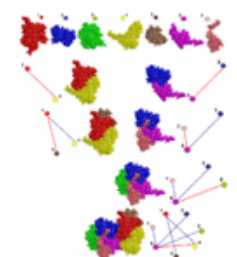


Figure 1: The prediction of the native quaternary structure of a protein complex. The diagram shows the predicted quaternary structure of a protein complex, which is a multi-molecular assembly. The subunits are represented by colored spheres, and the lines represent the predicted interactions between them. The diagram is divided into two parts: the top part shows individual subunits and their pairwise interactions, while the bottom part shows a more complex assembly of multiple subunits interacting together.

Short abstract: We have developed an algorithm that extends the application of docking to multimolecular assemblies. The algorithm should be particularly useful to predict the structures of large macromolecular assemblies, difficult to be determined experimentally. The algorithm well predicted quaternary structures of oligomers and multi-protein complexes, even for a structurally distorted input.

Full abstract: Abstract: The majority of proteins function when associated in multimolecular assemblies. Yet, prediction of the structures of multimolecular complexes has largely not been addressed, probably due to the magnitude of the combinatorial complexity of the problem. Docking applications have traditionally been used to predict pairwise interactions between molecules. We have developed an algorithm that extends the application of docking to multi-molecular assemblies. The algorithm should be particularly useful to predict the structures of large macromolecular assemblies, difficult to solve by experimental structure determination. The algorithm was applied to predict both quaternary structures of oligomers and multi-protein complexes. It produced an accurate predictions even on a structurally distorted input.

Introduction: Prediction of protein structures and their associations is essential for understanding of proteins function, regulation and cellular processes. Structure determination and high accuracy structure prediction schemes are major challenges of structural biology. Virtually all biological mechanisms involve a number of proteins that achieve their function by forming binary complexes, or in particular, multi-molecular assemblies. Yet, structure determination of protein assemblies is often harder than that of a single polypeptide chain, rising as the number of participating proteins and the size of the system increase. Thus, the development of prediction methods for multi-part or multi-molecular assemblies are extremely important. Protein-protein docking algorithms aim to predict the interactions between two given protein structure. Docking algorithms differ in the transformation search methods used to generate the candidate solutions and the scoring function applied to evaluate them. The scoring functions are still the weakest link of most docking algorithm. Thus, even when a near native complex is found it is not necessarily ranked as one of the top scoring solutions.

Multi-molecular assembly goal is to predict protein assemblies that involve more than two protein molecules. From the biological standpoint, prediction of multimolecular associations is undoubtedly extremely important. Yet, very little has been done toward this challenging application. The problem appears more difficult than the pairwise docking problem. Indeed, the problem is computationally difficult (NP-Hard). However, we show that the additional geometrical constraints embraced in the combinatorial problem may be exploited to increase the accuracy of the prediction and to outperform the pairwise interaction predictions that are involved in the target assembly. To the best of our knowledge CombDock, a combinatorial docking algorithm, is the first fully automated approach for the prediction of (hetero) multi-molecular assembly based only on structural models of their subunits.

Multi-molecular assembly via pairwise docking: The input to the multi-molecular assembly problem consists of a set of protein structural models. The goal is to predict the native complex formed by the interactions between the proteins. Geometrically, it is similar to the 3D puzzle reconstruction problem. In the multi-molecular assembly the “puzzle pieces” are the subunits, and the “puzzle solution” is the native complex. However, unlike a 3D puzzle where two connected pieces in the puzzle solution match perfectly, in the multi-molecular complex we would like to tolerate some extent of penetration due to the flexible nature of proteins. An additional difference lies in the

evaluation of the output complexes. In a puzzle we usually know the final structure of the complex, while in the molecular assembly problem the resolved complexes should be evaluated by the physical interactions. When solving a puzzle we usually apply a set of pairwise interactions, that gradually lead to the final solution. Here, we approach the multi-molecular assembly problem using a similar strategy. The algorithm has three stages: (i) all pairs docking, where candidate interactions between all pairs of the input structural models are predicted, (ii) combinatorial assembly, where complexes of the structural models are generated by applying different subsets of the candidate pairwise interactions, and (iii) final scoring where the complexes are evaluated and then clustered in order to discard redundant solutions.

Results and Conclusions:

Given the difficulties in experimentally determining multi-molecular complexes, the number of candidate target complexes with available structures in the protein data bank is limited. We have constructed a dataset of five different types of complexes, all with solved structures. The number of the subunits in the target complexes varies from 3 to 10. The sum of residues of the subunits in each complex varies from 328 in the smallest complex to 3519 in the largest one. Two of the cases are based on an unbound input, meaning that all or some of the subunits change their conformation when they dock to the complex. These cases are more difficult to predict. The algorithm has predicted a near native complex for each case, and ranked it in the top ten scoring ones. The results were encouragingly good even for those cases where the pairwise interactions between the subunits were not well predicted.

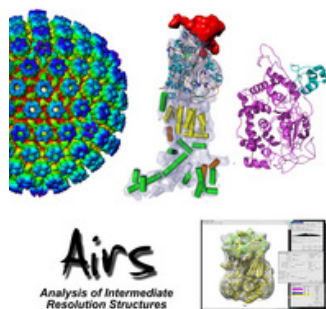
When we compare the quality of the prediction of the pairwise docking and the combinatorial docking, we see that the combinatorial approach outperforms the pairwise. The reason lies in the scoring function. In the more difficult cases, the leading docking algorithms succeed in finding the correct solutions but fail to highly rank them. The multiple approach attempts to combine the pairwise solutions to construct complexes that maximize the overall interactions between the subunits and at the same time, minimize the steric clashes. By doing so, it both excludes many of the candidate pairwise solutions that cannot be combined with others, and favors the set of pairwise solutions that may lead to fairly compact complexes. This set includes the near-native ones. We plan to generalize the CombDock algorithm to support multiple conformations for the same subunits, and to improve the scoring function that is used to evaluate the final and partial complexes. Such modifications are necessary for the algorithm to be a reliable tool for the prediction of multi-molecular complexes when the subunits are likely to show conformational changes upon binding. However, already this prototype version of CombDock has shown the advantages and some of the capabilities of the multiple docking approach, making us optimistic about the future role of such algorithms in the structural bioinformatics challenges.

Number: 77

Authors: Matthew L. Baker¹, Wah Chiu²

Baylor College of Medicine, ¹mbaker@bcm.tmc.edu ²wah@bcm.tmc.edu

Title: Analysis of Intermediate Resolution Structures



Short abstract: To facilitate the analysis of structures from electron cryomicroscopy, we have developed the Analysis of Intermediate Resolution Structures toolkit (AIRS). AIRS provides a mechanism to localize structures, recognize secondary structure, identify folds and model large macromolecular assemblies. This toolkit has been successfully applied to HSV-1 and RDV.

Full abstract: Electron cryomicroscopy can now routinely produce subnanometer resolution reconstructions for complex macromolecules. While electron cryomicroscopy has not yet reached atomic resolution, a wealth of structural and functional information can be mined at this resolution through integration of complementary techniques, such as modeling, bioinformatics, crystallography and structural analysis. To this end, we have developed AIRS, the Analysis of Intermediate Resolution Structures toolkit. The initial version of AIRS has incorporated several of the existing tools for structural analysis, including Helixhunter and Foldhunter, as well as several new algorithms for feature extraction and analysis. Through AIRS, secondary structure elements, protein folds and macromolecular architecture can be determined, ultimately leading to the construction of pseudo-atomic resolution structural models. Additionally, AIRS provides a platform for integrating ab initio and comparative modeling. AIRS is accessible through a command line interface as well as a graphical interface using UCSF's visualization package, Chimera. The graphical interface not only simplifies use but also provides additional functionality, such as rigid body refinement in Foldhunter, as well as detailed documentation. The tools within AIRS have been extensively tested on real and simulated data at various resolutions and are available with the EMAN image processing package at <http://ncmi.bcm.tmc.edu/~stevel/EMAN/>.

Of particular interest is the successful application of these tools to Rice Dwarf Virus (RDV) and Herpes Simplex Virus-1 (HSV-1) major capsid protein, VP5. In the 6.8 Å structure of Rice Dwarf Virus, both helices and sheets were identified in the two structural proteins. In conjunction with sequence analysis and fold recognition, pseudo-atomic models for these two RDV structural proteins were determined and later confirmed by the x-ray structure. For HSV-1 VP5, the x-ray structure for only one of the domains of the three domain major capsid protein was known. Again, using structural and sequence analysis, a pseudo-atomic model for the remainder of VP5 was determined, in which some structural similarity to lambdaoid bacteriophage structural proteins was observed in the floor domain of VP5. Additionally, ab initio modeling, where the cryoEM density was used as a constraint, allowed for the modeling of VP26, an accessory protein that binds only to hexonal VP5 molecules.

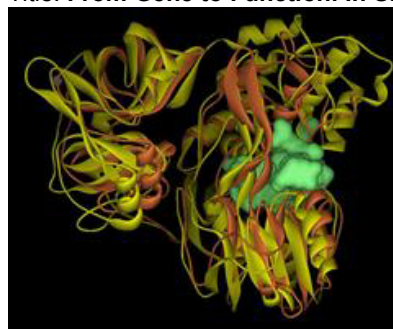
Laptop Session Abstracts

Number: 1

Authors: Anne Marie Quinn¹, Luke Fisher², Dana Haley-Vicente³

Accelrys, ¹aquinn@accelrys.com ²lfisher@accelrys.com ³danaHV@netscape.net

Title: From Gene to Function: In Silico Warfare on the West Nile Virus



Short abstract: Can we produce reliable structural models of the proteins of the West Nile virus even when sequence identity is low among homologs? Here we show that the DS GeneAtlas pipeline can be used to produce reliable structural and functional annotation of the proteins encoded by this genome.

Full abstract: ABSTRACT: Can we produce reliable structural models of the proteins encoded by the West Nile virus genome even when sequence identity is low among homologs? Such structural information is critical to further efforts to design drugs to fight the onslaught of disease. Here we show that the Discovery Studio® (DS) GeneAtlas pipeline can be used to produce reliable structural and functional annotation of the proteins encoded by the West Nile virus genome.

INTRODUCTION: The West Nile Virus is a single stranded RNA genome that belongs to the family of flaviviruses. It has a single open reading frame (ORF) encoding a polypeptide that is cleaved by co- and post-translational processing to produce structural and non-structural proteins. To date, no crystal or NMR structures have been produced for any of the protein products although a low-resolution cryo-electron microscopy structure was determined for the virus.

Structural information can be used to identify potential binding sites and classify protein function critical for developing new drug targets to combat the virus. The use of 3D homology models in the absence of an atomic resolution structure has been shown to be an effective alternative to study the structure and search for new lead candidates using structure-based drug design (SBDD). A recent publication shows several successful computational experiments that differentiate known inhibitors from a set of random compounds in a high-throughput docking procedure. Another recent publication employs the DS GeneAtlas pipeline to annotate the SARS virus in preparation for SBDD. Here we apply a similar approach to the West Nile virus proteome in hopes to functionally annotate the proteins and determine potential binding sites.

METHODS: DS GeneAtlas is an automated high throughput computational pipeline for structural and functional protein annotation. This program is part of Accelrys' DS Modeling software (http://www.accelrys.com/dstudio/ds_modeling/) for modeling and simulations. The DS GeneAtlas pipeline features domain prediction using TransMem and HMMer/Pfam, sequence similarity search using PSI-BLAST, secondary structure prediction using DSC, fold recognition using SeqFold, homology modeling using MODELER, and 3D annotation using binding site searching and 3D-motif predictions.

Twelve protein products derived from the ORF encoding the polyprotein sequence of the West Nile virus genome (accession number NC_001563) were selected for DS GeneAtlas analysis. These include the capsid protein (C), the membrane protein (M), the envelope protein (E), NS1, NS2a, NS2B, NS3, NS4A, NS4B, NS5 given here in the order that they are encoded by the ORF beginning at the 5' end. The precapsid (preC), or anchored C, includes the C-terminal signal sequence for the M-protein. PreM is the mature M protein with the aforementioned signal sequence cleaved. M is the fully processed protein. We identified the signalase and dibasic viral protease cleavage sites using DS Gene software. The cleavage site residues were taken from Chambers et. al.. A single

FastA file with the twelve processed protein sequences was the input for the DS GeneAtlas pipeline. SUMMARY OF RESULTS AND DISCUSSION: Reliable structural and functional annotations were generated for the envelope (E) protein, NS1, NS3, and NS5. No structural information was found for the cleaved M protein. In fact, the only annotation for the M protein was the prediction of a transmembrane domain from position 48 to 70 of the amino acid sequence. The uncleaved M protein, the capsid and precapsid proteins, and NS2a, NS2b, NS4a and NS4b had only very weak structural similarities to the PDB template proteins. The capsid protein was assigned a single transmembrane domain from position 44 to 66. Approximately the same region, position 44 to 58, was predicted to be helical by secondary structure analysis.

Predictably, the best template identified for the E protein was the crystalline structure of the dengue E protein (PDB template 1oam). At the sequence level, the dengue and West Nile proteins are 46% identical and 64% similar. This model spans the length of the West Nile protein, from position 3 to 400, excluding the last 101 amino acids of the C-terminus. This structure is in complex with N-Octyl-Beta-D-Glucoside, lending the potential for homology-based docking experiments on the West Nile E protein, as many of the residues in one of the four binding pockets are conserved. Further annotation by HMMer indicates an immunoglobulin-like domain from position 291 to 409 in the protein sequence. This domain is almost completely contained within the model. Two transmembrane domains were predicted in the C-terminus from position 452 to 474 and position 481 to 499. In addition, a conserved 3D-motif corresponding to a serine endopeptidases was predicted by the common structural clique method.

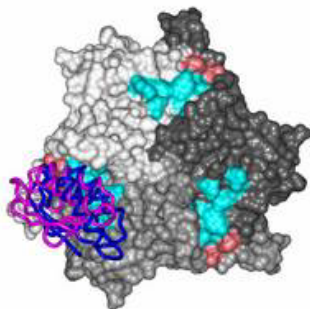
The best homology model for NS3 is based on the A chain of the template pdb structure 1cu1 (FIGURE). This template is a hydrolase complex from the Hepatitis C virus. The model identified five potential binding sites by the cavity search method. The largest binding site has a volume of 893 Ang.³ and may correspond to the ligand binding site in the 1cu1 template and model. The alpha-carbon superimposition of the model and template is 3.64 Ang. over a 536 residues out of 619 total in NS3. The PMF model score, 1.00, indicates an excellent fold. The Verify model score, 0.3 is good, but could have been even better if the sequence had been modeled in the presence of the B chain. The second best annotation was generated by the SeqFold fold recognition analysis and indicates a similar secondary structure prediction and good alignment as well as a larger coverage across the length of the sequence. Notably, this result demonstrates the ability to predict a reliable model structure even when sequence identity is a only 19%. The non-structural hydrophobic NS4a and NS4b proteins did not result in any homology model annotation. However, three and four transmembrane domains were reported by TransMem for NS4a and NS4b respectively.

Number: 2

Authors: Efrat Ben-Zeev¹, Miriam Eisenstein²

¹Weizmann Institute of Science, efrat.ben-zeev@weizmann.ac.il ²Weizmann Institute of Science, miriam.eisenstein@weizmann.ac.il

Title: Weighted Geometric Docking with MolFit: Incorporating External Information in the Rotation-Translation Scan



Short abstract: Weighted-geometric docking incorporates external data from different sources, such as biochemical and biophysical experiments and bioinformatics analyses, in the docking rotation-translation scan. The method is successful even when the weighted portion of the surface corresponded only partially and approximately to the binding site. Implemented in our program MolFit.

Full abstract: Weighted geometric docking is a prediction algorithm that matches weighted molecular surfaces. Each molecule is represented by a grid of complex numbers, storing information regarding the shape of the molecule in the real part and weight information in the imaginary part. The weights are based on experimental biochemical and biophysical data or on theoretical analyses of amino acid conservation or correlation patterns in multiple sequence alignments of homologous proteins. Only a few surface residues on either one or both molecules are weighted. In contrast to methods that use post-scan filtering based on biochemical information, our method incorporates the external data in the rotation-translation search, producing a different set of docking solutions, biased toward solutions in which the up-weighted residues are at the interface. Similarly, interactions involving specified residues can be impeded. Several tests indicated that the weighted-geometric algorithm produces much better ranking of the nearly correct prediction, and higher statistical significance. The method was found beneficial even when the weighted portion of the surface corresponded only partially and approximately to the binding site (1). Next, we successfully applied the method in the Critical Assessment of PRediction of

Interactions (CAPRI) experiment, using weights based on bioinformatics analyses (2). Another example was docking of colicin E3 to the 30S ribosomal subunit, in which the weights were based on biochemical data. Our model helped to predict the role of the catalytic residues of colicin E3 and the mechanism of inhibition by Immunity Protein (3).

1) Efrat Ben-Zeev and Miriam Eisenstein (2003). Prot: Structure, Function and Genetics, 2003 Jul 1;52(1):24-7. "Weighted Geometric Docking: Incorporating External Information in the Rotation-Translation Scan".

2) Efrat Ben-Zeev, Alexander Berchanski, Alexander Heifetz, Boaz Shapira and Miriam Eisenstein (2003). Proteins: Structure, Function and Genetics, 2003 Jul 1;52(1):41-6.

"Prediction of the unknown: Inspiring experience with the CAPRI Experiment".

3) Efrat Ben-Zeev, Raz Zarivach, Menachem Shoham, Ada Yonath and Miriam Eisenstein (2003). J. of Biomolecular Structure and Dynamics, 2003 Apr;20(5):669-76.

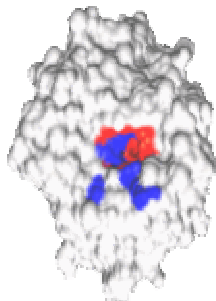
"Prediction of the Structure of the Complex Between the 30S Ribosomal Subunit and Colicin E3 via Weighted-Geometric Docking".

Number: 3

Authors: **James Bradford**¹, **David Westhead**²

¹University of Leeds, School of Biochemistry and Molecular Biology, bmbjrb@bmb.leeds.ac.uk ²University of Leeds, School of Biochemistry and Mi, westhead@bmb.leeds.ac.uk

Title: **Prediction of Protein-Protein Binding Sites using Support Vector Machines**



Short abstract: We have applied a machine-learning approach, the support vector machine, to the prediction of protein-protein binding sites. Our method classifies surface patches as part of or outside a binding site, and has a high success rate with broad specificity, being able to predict both transient and obligomeric binding sites.

Full abstract: We have applied an increasingly popular machine-learning approach, the support vector machine (SVM), to the prediction of protein-protein binding sites. We trained the SVM to distinguish between interacting and non-interacting surface patches using six properties: interface residue propensity, hydrophobicity, sequence conservation, surface topography, solvent accessibility and electrostatic potential, and then used this SVM as a component of our prediction method where it was applied to surface patches centred on every surface atom. The final interaction site prediction was made using a novel process of patch overlap analysis and ranking, based on the SVM's output of confidence values for patches predicted to be interacting. Using a leave-one-out cross validation procedure, we were able to successfully predict the location of the binding site on 76% of the 180 proteins in our own data set - the largest manually produced data set of its kind. Importantly, prediction success was not restricted to a single complex type thus our method is applicable to predicting both transient and obligate interfaces. With heterogeneous cross-validation, where we trained the SVM on transient complexes to predict on obligate complexes (and vice versa), we still achieved comparable success rates to the leave-one-out cross validation suggesting that sufficient properties are shared between transient and obligate interfaces. Furthermore, upon closer inspection of six "failures", we found that the predicted patches formed part of another binding site suggesting that our method is identifying important functional sites even if the anticipated binding site proves elusive.

Number: 4

Authors: **Kim Henrick**¹, **Tom Oldfield**², **Adel Golovin**, **John Tate**, **Sameer Velankar**, **Harry Boutselakis**, **Dimitris Dimitropoulos**, **Peter Keller**, **Eugene Krissinel**, **Phil McNeil**, **Jorge Pineda**, **Abdelkrim Rachedi**, **Antonio Suarez-Uruena**, **Jawahar Swaminathan**, **Mohamed Tagari**

European Bioinformatics Institute, ¹henrick@ebi.ac.uk ²oldfield@ebi.ac.uk

Title: **MSD Relational Database, Search and Visualisation of Queries and Results**



Short abstract: The MSD macromolecular structure database together with search interfaces will be described as an integrated system that includes structure matching through to active site visualisation.

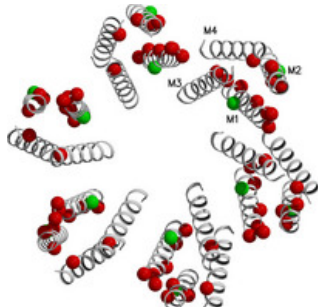
Full abstract: The E-MSD macromolecular structure relational database(<http://www.ebi.ac.uk/msd>) is designed to be a single access point for protein and nucleic acid structures and related information. The database is derived from Protein Data Bank (PDB) entries. Relational database technologies are used in a comprehensive cleaning procedure to ensure data uniformity across the whole archive. The search database contains an extensive set of derived properties, goodness-of-fit indicators, and links to other EBI databases including InterPro, GO, and SWISS-PROT, together with links to SCOP, CATH, PFAM and PROSITE. A generic search interface is available, coupled with a fast secondary structure domain search tool are aspects of our continuous process of enhancing the quality and consistency of macromolecular structure data working towards the integration of various bioinformatics data resources. New simple form-based interfaces that allows users to query the MSD directly and the MSD atlas pages show all of the information in the MSD for a particular PDB entry. Additional search interfaces aimed at specific areas of interest, such as the environment of ligands and the secondary structures of proteins have been released. We have also implemented a novel search interface that begins to integrate separate MSD search services in a single graphical tool. We have worked closely with collaborators to build a new visualization tool that can present both structure and sequence data in a unified interface, and this data viewer is now used throughout the MSD services for the visualization and presentation of search results.

Number: 6

Authors: Sarel J. Fleishman¹, Vinzenz M. Unger, Mark Yeager, Nir Ben-Tal

¹Tel-Aviv University, sarel@post.tau.ac.il

Title: A Model Structure of the Gap-Junction Transmembrane Domain Specifying C-alpha Positions



Short abstract: A set of methods for predicting the orientations of transmembrane alpha helices was developed and applied to predicting the gap junction's structure. The model provides a structural basis for understanding the different physiological effects of almost 30 mutations and polymorphisms, revealing an intimate relationship between molecular structure and disease.

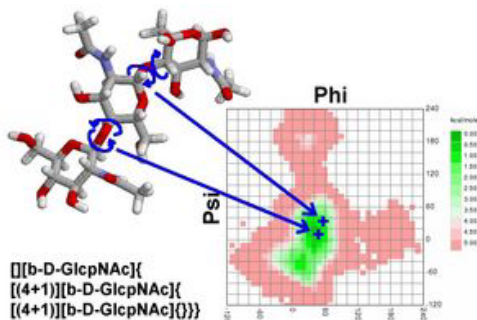
Full abstract: We have developed a set of novel computational methods for predicting the orientations of helices in transmembrane proteins based on low-resolution data (5-10 Å). These methods served to predict the structure of the transmembrane domain of the gap junction channel. Gap-junction channels connect the cytoplasm of apposed cells, and are comprised of a hexamer of connexin proteins on each of the two cell membranes. Using a new three-dimensional cryo-electron microscopy (cryo-EM) map of gap-junction channels at 5.7 Å in-plane resolution, we identified the positions and orientations of the transmembrane helices within each membrane. The four hydrophobic domains in connexin sequences, termed M1-M4, were assigned to the helices in the cryo-EM map based on biochemical and phylogenetic data. Evolutionary conservation and an analysis of compensatory mutations in connexin evolution were used to identify the packing interfaces between the helices. The final model, which specifies the coordinates of C-alpha atoms in the transmembrane domain, provides a structural basis for understanding the different physiological effects of almost 30 mutations and polymorphisms, and reveals an intimate relationship between molecular structure and disease (see Figure). The model presents a framework for testing various hypotheses regarding the stability and conformational changes underlying gap-junction function.

Number: 7

Authors: Thomas Lutteke¹, Claus W. von der Lieth²

DKFZ Heidelberg, Central Spectroscopic Department, ¹t.luteteke@dkfz.de ²w.vonderlieth@dkfz.de

Title: Towards Deciphering the Structural Features of Glycosylation Sites: Analysis of Carbohydrate-Related Data contained in the Protein Data Bank



Short abstract: www.glycosciences.de provides several tools to examine carbohydrate 3D structures derived from PDB entries using `pdb2linucs`: GlyTorsion statistically analyses various torsion angles. GlySeq generates statistics on amino acids around glycosylation sites. Carp performs a Ramachandran-like plot of carbohydrate linkage torsions. Data sets are updated weekly with new PDB entries.

Full abstract: Protein glycosylation is probably the most common and complex type of co- and posttranslational modifications encountered in proteins. Glycosylation patterns are sensitive markers of changes within the environment of cells. Glycan profiling of normal and diseased forms of a glycoprotein has provided new insights for future research in rheumatoid arthritis, prion disease and congenital disorders of glycosylation [1,2]. The oligosaccharide moieties cover a range of diverse biological functions, several of which depend on the precise three-dimensional shape of the glycan. Therefore knowledge of the 3D-structure of the glycan is a prerequisite for a full understanding of the biological processes glycoproteins are involved in [3,4]. The development of approaches having explanatory power to reliably predict e.g. occupied glycosylation sites or carbohydrate binding sites is hampered by the fact, that only very little experimental data are available. To derive rules for the properties of glycosylation sites or carbohydrate binding proteins, large data sets are necessary. A source for such data is the Protein Data Bank (PDB) [5]. To receive the data from `pdb`-files, some new functions were implemented into the carbohydrate detection software `pdb2linucs` (<http://www.glycosciences.de/tools/pdb2linucs/>) [4]. The program stores torsion angles, glycoprotein sequences and amino acids in the vicinity of carbohydrates in `xml`-files, from which they can be analysed by further programs. Data sets are updated weekly with new PDB entries. Examination of torsion angles is performed by the program GlyTorsion, which can be accessed over the internet (<http://www.glycosciences.de/tools/glytorsion/>). Previous publications reflect a current state only and are limited to linkages between a small set of residue types [3,6]. In contrast, the web interface provides access to a regularly updated data set and allows the user to specify the residues of his interest. Data derived from GlyTorsion could be used to proof computationally generated conformation maps or to improve methods to calculate carbohydrate structures, e.g. in silico glycosylation of proteins [7,8]. Another application of linkage torsion data extracted by GlyTorsion is to serve as a scattered background in Ramachandran-like plots for specific carbohydrate linkages. This option is implemented in carp (CArbohydrate Ramachandran Plot, <http://www.glycosciences.de/tools/carp/>). This tool searches an uploaded `pdb`-file for carbohydrates using the `pdb2linucs` algorithm [4], generates a plot for each linkage type found in the uploaded structure and displays as scattered background all other torsion angles contained in PDB for this specific linkage. Alternatively, force-field based conformational maps automatically generated from long time high temperature molecular dynamics simulations, which are stored in the GlycoMapsDB (<http://www.dkfz.de/spec/glycomaps/>), can be selected as plot background.

The software GlySeq (<http://www.glycosciences.de/tools/glyseq/>) is intended for analysis of amino acid sequences of glycosylated proteins. While the absence of a glycan in the `pdb`-file does not necessarily mean that a potential glycosylation site is unoccupied in the native protein, its presence provides evidence for the occupancy of a glycosylation site [9]. Therefore, glycosylated protein structures in the PDB yield sufficient data for sequence analysis around glycosylation sites. Before starting the analysis, redundant sequences were removed from the data set.

N- and O-glycosylation sites show significant differences in the amino acid composition of the surrounding residues. While aromatic residues are more frequent around N-glycans than within average sequences, they occur seldom near O-glycans. The latter ones show an increase in Ser/Thr residues, which often are glycosylated themselves. Furthermore, O-glycans are much more often located near the beginning or the end of a sequence than N-glycans. At N-glycosylation sites, Pro occurs seldom at positions between -2 and +3, but with increased frequency at positions around this area. Near O-glycans, Pro is very frequent, especially at positions with an odd distance to the glycosylation site. With the perpetual growth of the PDB the number of glycoprotein sequences available for statistical examination will also grow steadily. The use of this data as test or training sets for glycosylation prediction software like NetNGlyc [10] might help to increase the accuracy of predictions.

We also intend to analyse the amino acid composition within the spatial vicinity of carbohydrate residues found in the PDB. Since that investigation can not only be done on covalently attached glycans but also on non-covalently bound ligands, it might lead to a better understanding of the processes included in protein-carbohydrate interactions.

References:

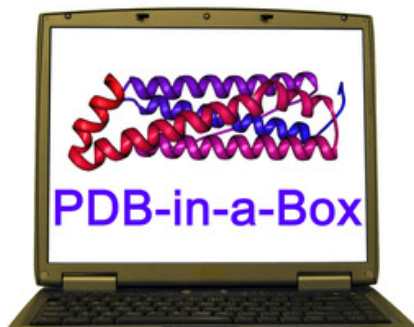
1. Peracaula R, Royle L, Tabares G, Mallorqui-Fernandez G, Barrabes S, Harvey DJ, Dwek RA, Rudd PM, De Llorens R. *Glycobiology*, 2003, 13, 227-44.
2. Butler M, Quelhas D, Critchley AJ, Carchon H, Hebestreit HF, Hibbert RG, Vilarinho L, Teles E, Matthijs G, Schollen E, Argibay P, Harvey DJ, Dwek RA, Jaeken J, Rudd PM. *Glycobiology*, 2003, 13, 601-22.
3. Wormald MR, Petrescu AJ, Pao YL, Glithero A, Elliot T, Dwek RA. *Chem Rev*, 2002, 102, 371-386.
4. Lutteke T, Frank M, von der Lieth CW. *Carbohydr Res*, 2004, 339, 1015-1020.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. *Nucl Acids Res*, 2000, 28, 235-242.
6. Imberty A, Perez S. *Protein Eng*, 1995, 8, 699-709.
7. Bohne A, von der Lieth CW: *Pac Symp Biocomput*. 2002;:285-296.
8. Frank M, Bohne-Lang A, von der Lieth CW: *In Silico Biol*. 2002;2(3):427-39.
9. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR. *Glycobiology*, 2004, 14, 103-114.
10. Gupta R, Brunak S. *Pac. Symp. Biocomput*. 2002, 310-22.

Number: 8

Authors: **Wolfgang F. Bluhm¹** on behalf of the PDB team

¹RCSB Protein Data Bank, wbluhm@sdsc.edu

Title: **PDB-in-a-Box**



Short abstract: The Protein Data Bank (<http://www.pdb.org/>, PDB), maintained by the Research Collaboratory for Structural Biology (RCSB), is completely reengineering its Web site and database. We will demonstrate one product of this reengineering: a stand-alone version of the entire Web site and database on a laptop computer.

Full abstract: The Protein Data Bank (<http://www.pdb.org/>, PDB), which is maintained by the Research Collaboratory for Structural Bioinformatics (RCSB), is in the process of reengineering the current RCSB PDB Web site and database. One of the new concepts of the reengineered RCSB PDB is the ability to run a stand-alone version of the entire Web site and database on a laptop computer. We will demonstrate a fully functional version of this "PDB-in-a-box".

The general requirements are fairly moderate and met by most modern laptops. The basic application requires only a Java Application Server (such as JBoss), and a SQL-92 compliant relational database such as MySQL. Minimal hardware requirements are 512 MB RAM and 20 to 40 GB of free hard disk space. A variety of operating systems can be used, such as Windows, Linux, and Mac OS X. Some graphical components require the Java 3D plugin.

PDB-in-a-box provides the same basic functionality as the reengineered RCSB PDB Web site hosted by the main servers of the RCSB PDB at the San Diego Supercomputer Center. Both the new site and PDB-in-a-box are based on the curated data that has resulted from the RCSB PDB data uniformity project (see <http://www.rcsb.org/pdb/uniformity>). The database is loaded from XML-formatted versions of the curated data files, which have been publicly available since June 2003 at <ftp://beta.rcsb.org/pub/pdb/uniformity/data/XML/>. The reengineered RCSB PDB Web site has been designed with novice to experienced users in mind and driven primarily by user feedback. Feedback has come from the RCSB PDB help desk (info@rcsb.org), discussions at conferences, one-on-one focus group sessions with select users, and the input of an expert usability engineer. The site has been implemented as a three-tier architecture which provides a more robust, modular and more easily maintainable system.

New features of the reengineered RCSB PDB include:

- 1) Better navigation and search tools for both structure and Web site content.
- 2) Browsing of structures classified according to taxonomy, structural features, functional features, known biochemistry and relationship to specific diseases.

- 3) Searching of PubMed abstracts which represent the primary structural citations.
- 4) Detailed reports for single structures and sets of multiple structures with additional experimental details.
- 5) New visualization options.
- 6) Data access through Web Services.

The PDB's core mission remains the distribution of and access to the primary experimental data from three-dimensional structural studies on biological macromolecules. The reengineered RCSB PDB site has the added ability to address a broader range of biological questions, for example:

- 1) What is the role of ferritin in iron metabolism?
- 2) What viral capsid proteins are in the PDB?
- 3) What is the role of enzymes in human apoptosis?

This abstract was submitted on behalf of the entire RCSB PDB team.

A complete list of current RCSB PDB staff members is available at <http://www.rcsb.org/pdb/rcsb-group.html>.

The RCSB PDB is supported by funds from the National Science Foundation, the Department of Energy, and the National Institutes of Health.

Number: 10

Authors: **Howard J. Feldman**¹, **Christopher WV Hogue**², **Kevin Snyder**

The Blueprint Initiative, ¹feldman@mshri.on.ca ²hogue@mshri.on.ca

Title: **MMDBBIND - Macromolecular Interactions with Atomic Level Detail**



Short abstract: MMDBBIND provides over 18,000 high quality fully annotated, searchable, atomic-level detail molecular interaction records to BIND (Biomolecular Interaction Network Database). Interactions can be viewed with the Cn3D structure viewing software, with interacting residues highlighted. These could be used, for example, to generate empirical protein-protein docking potentials.

Full abstract: The Biomolecular Interaction Network Database (BIND) (<http://www.bind.ca>) is an expanding database of biomolecular interaction, pathway and complex information. All information is stored in database records that are freely available through a web interface that allows users to query, view, and submit records. We have treated the Protein Data Bank (PDB) as a source for molecular interaction data and produced MMDBBIND, a set of over 18,000 algorithmically generated fully annotated BIND interaction records. Initially, all pairwise protein-protein, protein-DNA, protein-RNA, DNA-DNA, DNA-RNA and RNA-RNA interactions contained within each PDB file were recorded, with both residue-level and atomic-level detail at the interaction site. A minimum of two residue pairs were required to form an interaction. Filters were applied to remove most crystal packing artifacts using EBI's PQS system, and to remove double-stranded DNA. Redundant interaction interface surfaces have been collapsed into single records in the following manner. Using the nrPDB table provided by NCBI, we group together interactions obtained from PDB files (including itself) within the same NCBI non-redundant group (BLAST p-value cutoff 10e-40). Within each group, an N-by-N interaction comparison is carried out. Interactions determined to be redundant are clustered together, and a representative is chosen. A single BIND record is generated for each redundant group, with the annotation coming from the representative. Two interactions are considered to be redundant if their corresponding sequences have greater than 95% sequence identity AND at least 90% of the binding site residues of one interaction map via the sequence alignment to the binding site residues of the second interaction.

Records have additionally been annotated with the same level of detail as a hand-curated BIND record. Experimental descriptions have been parsed from PDB comments and translated into structured English sentences. Where available, annotation and short labels have also been added using LocusLink, SGD and other sources. Links to PDB, MMDB (NCBI's ASN.1 mirror of the PDB) and MEDLINE are also provided for each record. The BIND interface allows a user to view the 3D structure of each MMDBBIND record in Cn3D (<http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>) with the interacting residues on each chain highlighted. Raw BIND data has also been made available on Blueprint's FTP site (<ftp.blueprint.org>). This consists of two flat-files – one providing the sequence pairs in each interaction with binding sites represented in uppercase, and one listing binding site residue pairings for each interaction. The process will be run on a monthly basis, in the near future, to add interactions for newly deposited PDB structures.

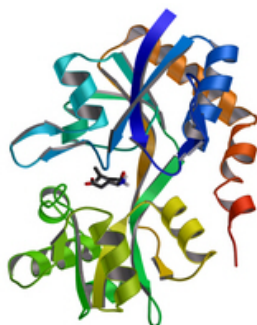
These computer-generated BIND records will offer biologists a tool to visualize the interaction information of PDB between discrete molecules at the atomic level of detail. Additionally, they provide a unique way of grouping together redundant protein-protein interactions. Since atomic level detail is given in the records, they will also serve as an ideal source for collecting statistics on protein binding site properties. Such statistics can be utilized, for example, to develop empirical scoring functions for use in protein-protein docking.

Number: 11

Authors: Philip C. Biggin¹, Yalini Arinaminpathy², Mark SP Sansom

Oxford University, ¹phil@biop.ox.ac.uk ²pathy@biop.ox.ac.uk

Title: Conformational Dynamics of Glutamate Receptors: Simulation Studies.



Short abstract: The precise mechanism of partial agonism in ionotropic glutamate receptors remains unclear. We present results from molecular dynamics simulations of the ligand-binding core (S1S2) which suggest how partial agonists can induce single-channel currents of the same magnitude as full agonist but with a reduced frequency as recently observed experimentally.

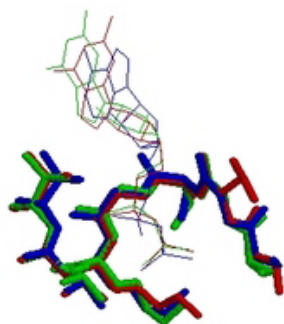
Full abstract: Ionotropic glutamate receptors are central to fast synaptic neurotransmission in the brain. A series of x-ray structures for the ligand-binding (S1S2) domain of the GluR2 (-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-sensitive receptor have suggested how the degree of closure of this domain appears to correlate with the agonist efficacy. More recent single channel experiments revealed that partial agonist could open the channel to similar conductance levels, but with reduced frequencies, as full agonists do. In addition, a series of x-ray structures for a similar region from the ligand-binding core of the prokaryotic glutamate receptor, GluR0 demonstrated how a partial agonist could induce a degree of domain closure similar to that found for a full agonist. Thus, the precise mechanism of partial agonism is not well understood. We have performed molecular dynamics simulations (over 100ns in total) to explore the dynamics of this domain from the two different proteins. Simulations have been performed based on these crystal structures with the natural agonist glutamate, the partial agonist kainate (in GluR2), the partial agonist serine (in GluR0) and two conformations of the apo state (closed and open). For both GluR2 and GluR0 our results show that the degree of motion of one sub-domain compared to the other is dependent on the agonist present, supporting previous observations. In the case of the GluR2 simulations we made two key observations; firstly, the open-apo simulation exhibited significant inter-domain motions such that it closed to a state resembling the glutamate-bound crystal structure, thus defining an open to closed pathway, and secondly, we observed some closure for the kainate bound structure. The x-ray structure suggested that this should not be possible due to steric clashes of the isoprenyl group. Upon closer inspection of this simulation, we found that the kainate had adopted a change in conformation and orientation within the binding site that allowed the D1 and D2 lobes to close together further than observed crystallographically. This result provides an alternative resolution to the question of how kainate may behave as a partial agonist at the macroscopic level and yet gives subconductance levels identical to full agonists. Finally, we observe an opening event in one of the GluR0 simulations, thus defining a closed to open pathway.

Number: 12

Authors: Nicola D. Gold¹, Richard M. Jackson²

¹University of Leeds, n.d.gold@leeds.ac.uk ²jackson@bmb.leeds.ac.uk

Title: A Searchable Database for Comparing Protein-Ligand Binding Sites for the Discovery of Structure-Function Relationships



Short abstract: It may be possible to infer common functional roles for proteins with similar ligand binding sites. We have therefore developed a web accessible database of ligand binding sites combined with a similarity searching method based on geometric hashing to identify binding sites with similar atomic arrangements and properties.

Full abstract: Structural classification databases have been used extensively to assist function prediction of new and unknown proteins by similarity in their folds (SCOP, CATH, FSSP). While this approach is useful it is known that a single protein fold can provide many different functions. In these cases additional functional information can often be provided by comparing the detail of protein structure at ligand binding sites and enzyme active sites to known and characterised sites in the database (e.g. TESS, ASSAM, PINTS).

Structural genomics projects are increasingly solving structures of hypothetical proteins of unknown function and there is a need to predict protein function from structure. Our aim is to aid the characterisation of these proteins by structure based prediction of ligand binding site similarity. We have developed a large database of ligand binding sites extracted automatically from the Macromolecular Structure Database. This has been combined with a new method (Brakoulias and Jackson, 2004) based on geometric hashing to assess the extent of their similarity. In particular the binding sites of the large class of nucleotide ligands (e.g. ATP/ADP, GTP/GDP, FAD, NAD) is studied here but will be expanded to include other small-molecule ligands in the future.

Similarity in the spatial arrangements of atoms between any two sites might indicate that they bind similar ligands and thus may exhibit similar functions. The geometric hashing method provides a score of similarity, an RMSD and a superposition of equivalenced atoms for each pair of compared binding sites. The similarity score allows us to identify the most similar ligand binding sites to a defined query from our database and may allow common function to be inferred. These preprocessed data are stored in a World Wide Web accessible database which is searchable with a PDB code and ligand information (such as ligand name, number and chain). Submission of these data rapidly returns a ranked list of similar ligand binding sites (above a certain similarity score cut-off). Additional annotation is also provided such as the SCOP (Structural Classification Of Proteins) codes and each hit is coloured according to its similarity to the overall family and fold of the query protein. Optimal superpositions of the binding site and ligands of interest can then be performed allowing further examination and visualisation with molecular graphics. A multiple alignment of structurally equivalenced atoms is also provided. Our current focus is to define similarity with a statistical E-value using the method of Stark and coworkers (Stark et al., 2003) in order to give a measure of confidence in the identified similarity.

The method is successful in identifying close family and superfamily relationships prior to their classification in the SCOP and CATH databases and has also been applied successfully to identifying more distant evolutionary relationships at the fold level or even beyond. In addition we are able to detect similarity between binding sites where the ligands are similar but not identical.

References

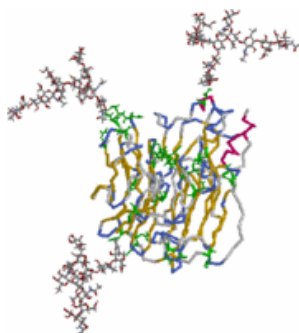
- Brakoulias A., Jackson RM. "Towards a Structural Classification of Phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching" *Proteins; Structure function and bioinformatics*, in press.
- Stark A, Sunyaev S and Russell RB. (2003) "A model for statistical significance of local similarities in structure" *JMB*, **326** 1307-1316.

Number: 13

Authors: Andreas Bohne-Lang¹, Claus W. von der Lieth

¹German Cancer Research Center, a.bohne@dkfz-heidelberg.de

Title: Glyprot: a Web-tool for in-silico Glycosylation of Proteins



Short abstract: Over 50% of the proteins are glycosylated. But many of the 3d structures of proteins stored in PDB have no or only a small carbohydrate part. However, our free web-tool is capable to create a 3d protein structure with attached n-glycan — in-silico.

Glyprot: <http://www.dkfz.de/spec/glyprot/>

Full abstract: Introduction:

Animal cell surfaces are rich in glycoconjugates and over 50% of the proteins are glycosylated [1]. These glycans are placed on the proteins which are often part of the plasma membrane which creates a meshwork the projects up to 100Å from the cell. This meshwork called glycocalyx represents the outermost surface and plays a defining role in interactions (e.g. cell-cell interactions, bacteria-cell, or virus-cell interactions). A major role of membrane protein glycosylation is to present various terminal structures that can be bounded by lectins – a carbohydrate binding protein. N-glycans are complex carbohydrates, which are linked to the protein by asparagin (N).

Unfortunately, many of the 3d structures of proteins stored in PDB have none or only a small carbohydrate part. The reason for this is that carbohydrates are too flexible for x-ray structure analysis and many of the proteins are created recombinant without the posttranslational modifications.

The Glyprot web-tool closes this gap by offering a service where the user can create a 3d model of a protein with its attached n-glycans.

Methods:

Beginning with the 3d model of the protein taken from PDB database [2], the Glyprot tool analyses the protein. First, the aminoacid sequence is detected and potential glycosylation sites are extracted. A glycosylation site for n-glycans is characterized by a special amino acid sequence. Whenever the pattern NxS/T (with x being every amino acid except for proline) is found, it will be marked. In a second step the program looks for linked carbohydrates in the 3d model of the protein's crystal structure. If a linked saccharide is found, the torsion angles are measured and stored. The program checks in a third step the potential n-glycan glycosylation sites if enough space for a n-glycan is available. Normally, potential n-glycan glycosylation sites are placed at the outer area of a protein; thus, normally a geometric solution is possible. For protein modelers this tool can be quite useful because it diagnoses the correct placement and orientation of the glycosylation site. The probing angle combinations are taken from a systematic analysis of the angle setting in the crystal structure [3,4]. The probing angle combinations are analysed by a calculation of the angle setting in the crystal structure. In a last step the user can select a n-glycan structure and add it to the glycosylation site.

There are three ways to select a n-glycan for in-silico glycosylation of the protein:

1. One typical representative of the common n-glycan classes can be selected directly from a pull-down menu.
2. A database with over 1000 3d n-glycan structures was built up and connected to the Glyprot tool. This n-glycan database stores MM3 force field optimized 3d structures of the glycans created with the carbohydrate builder tool Sweet2 [5]. The user can search by using different options (e.g. mass values, structure description, or monosaccharides).
3. Unless the user finds his structure in the database, it is possible to create a new n-glycan structure by Sweet2. The program suggests the geometrical angle values and whenever a n-glycan part is found in the 3d structure of the protein, this value is also offered. The user can change the angle settings by editing the web form field for angle values.

Results:

After the settings of the parameters are completed, the user can create the 3d model of an in-silico glycosylated protein online and get back the 3d structure encoded in PDB format. Additionally, Oleg Tsodikov's Surface Racer [6] calculates structure properties like solvent accessible surface areas or molecular surface areas. The user can display the 3d model by the Java applet WebMol [7] when the Chime plug-in is installed.

URL: [GlyProt Web Tool](http://www.dkfz.de/spec/glyprot/)

References:

1. Apweiler R, Hermjakob H, Sharon N.; On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database.; *Biochim Biophys Acta*. 1999; 1473(1): 4-8.; Review..
2. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, 2002; 28 pp. 235-242
3. Imberty A, Perez S.; Stereochemistry of the N-glycosylation sites in glycoproteins. *Protein Eng.*; 1995; 8(7): 699-709.

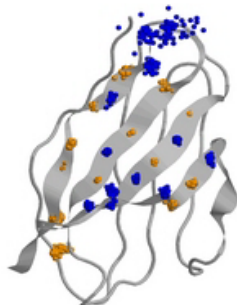
4. Petrescu AJ, Petrescu SM, Dwek RA, Wormald MR.; A statistical analysis of N- and O-glycan linkage conformations from crystallographic data. *Glycobiology*.; 1999; 9(4): 343–52
 5. Bohne A, Lang E, von der Lieth CW.; SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides.; *Bioinformatics*; 1999; 15(9): 767-8.
 6. Tsodikov, O. V., Record, M. T. Jr. and Sergeev, Y. V.; A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.*, 2002, 23, 600-609.
 7. Walther D.; WebMol--a Java-based PDB viewer.; *Trends Biochem Sci.* 1997; 22(7): 274–5.
-

Number: 14

Authors: **Vladimir Potapov¹, Vladimir Sobolev², Marvin Edelman, Alexander Kister, Israel Gelfand**

Weizmann Institute of Science, ¹vladimir.potapov@weizmann.ac.il ²vladimir.sobolev@weizmann.ac.il

Title: **Interface Cores in Sandwich-like Proteins**



Short abstract: We define protein-protein interface and domain cores for immunoglobulins containing about half the residues and surface areas of the full interfaces. Residues of the two cores are structurally connected, imparting rigidity at the interface core. The rule of positional connectivity extends generally to sandwich-like proteins interacting in a sheet-sheet fashion.

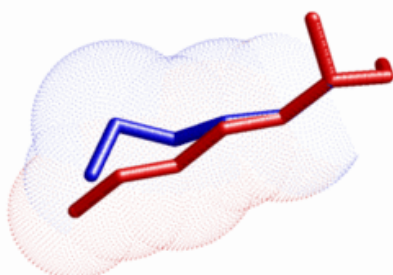
Full abstract: Prediction of a protein-protein interface is still restricted in accuracy. As a result, intensive efforts are being applied to extract additional levels of information from the database relating to surface-surface interaction. We find that where statistical data exist (as in the case of the resolved heavy and light chains of Ig proteins), subsets of interface positions can be extracted that define an interface core and a domain core, and offer a new tool with a reduced target for the analysis of sheet sheet interactions in sandwich-like proteins. The refined interface cores contain approximately half the residues and half the surface areas of the full interfaces. The techniques developed for the Ig interfaces and domain cores prove adequate to extract first-approximation cores for a set of non-Ig sandwich-like interfaces, thus extending the usefulness of the approach. Our analysis reveals that most of the positions in sandwich-like proteins crucial for sheet sheet inter-domain and intra-domain interactions are adjoined one to another in sequence. This finding was derived independent of geometric considerations, however beta-sheet side chain geometry clearly dictates such neighboring. Since the domain core is commonly accepted as the most stable part of the protein structure we can expect that adjoining residues will also be rather restricted in their flexibility. The tight clustering of the interface core positions across the Ig data set bears this out. Thus, the juxtaposition of interface and domain core residues experimentally supports the concept of a rigid substructure on the protein surface involved in complex formation.

Number: 16

Authors: **Eran Eyal¹, Vladimir Sobolev, Rafael Najmanovich, Brendan J. McConkey, Marvin Edelman**

¹Weizmann Institute of Science, Rehovot, Israel, eyale@wicc.weizmann.ac.il

Title: **Modeling Side Chain Conformations using Contact Surfaces and solvent Accessible Surface**



Short abstract: Contact surface area and chemical properties of atoms form the core of a scoring function to concurrently predict conformations of amino acid sidechains. The program (<http://sgedg.weizmann.ac.il/sccomp.html>) combines accuracy and speed. Most atoms prefer intramolecular surface contact over solvent contact. This might be the driving force for maximizing protein packing.

Full abstract: Abstract: Contact surface area and chemical properties of atoms are used to concurrently predict conformations of multiple amino acid side chains on a fixed protein backbone. The scoring function is particularly suitable for modeling partially-buried side chains. Both iterative and stochastic searching approaches are used for modeling multiple side chains. The programs have a good combined execution time and accuracy. We find that the differential between the atomic solvation parameters and the contact surface parameters (including those between non-complementary atoms) is positive; i.e., most protein atoms prefer surface contact with other protein atoms rather than with the solvent. This might correspond to the driving force for maximizing packing of the protein. The programs (Scomp-S and Scomp-I) can be accessed through the web (<http://sgedg.weizmann.ac.il/scomp.html>) and are available for several platforms.

Introduction: Side chain modeling is a necessary and important step in constructing complete structural models of proteins. Various approaches to construct structural models, such as comparative modeling, threading and ab initio prediction usually assign side chain location artificially as a sub-task of the entire modeling procedure(1). Modeling of side chains is also required for flexible molecular docking, for predicting local structural and stability changes upon mutations and for protein design.

It recently was suggested that the combinatorial search problem is not that severe in practice (2) and the focus of research has shifted towards the scoring function. More attention is now given to the solvation terms (3,4) and detailed rotamer libraries allow knowledge-based derivation of pseudo-energy values for intra-residue energy and local backbone interactions (5-7). An optimized scoring function that includes geometric characteristics of interactions gives impressive results and was demonstrated to perform significantly better than standard force fields (6).

We developed a new scoring function based on surface complementarity that considers geometric and chemical compatibility and solvent accessible surface. A function of this sort might be appropriate with only minor modification for both predicting structural changes and estimating stability of point mutations.

Results and discussion: A surface complementarity term based on contact surface area and chemical properties of the atoms, is the core of our scoring function to predict side chain conformations. A solvation term was incorporated based on the solvent contact surface area. An excluded volume term for clashing atoms and an intra-residue energy term based on rotamer probabilities were also included. The general form of the scoring function is: $E_{score} = E_{comp} + K_{vol}E_{vol} + K_{prob}E_{prob} + K_{sol}E_{sol}$

The exact form of each of the components can be seen in Eyal et al. (8). The relative contribution of various terms in the scoring function was calibrated by implementing a genetic algorithm. The root mean square deviation of side chain atoms between the model and native structures was minimized. Each fitness evaluation included modeling one of 2983 side chains in a training set of high resolution proteins, while all others were held fixed in the native conformation. Examination of the contribution of the various terms to the combined scoring function revealed that the intra-residue energy term significantly contributes to the prediction of exposed residues, while the surface based terms are important for prediction of partially buried side chains.

We also tried to optimize the value of the atomic solvation parameters (ASPs). Although there was no convergence to a single set of values, we noticed that all the ASPs had positive values almost always greater than 1 (the weight of contacts between chemically non-complementary atoms). In terms of contact surface area, inter-molecular interaction is therefore favorable over interaction with the solvent. This observation is in general agreement with other works that employed ASP models for structural modeling and found that all ASPs should be positive (9-10).

Two searching methods were applied with the scoring function to concurrently model multiple side chains: a simple iterative self-consistent one, and a stochastic method, based on the Boltzmann distribution. Our programs, with relatively fast execution times, correctly predict χ_1 angles for 92-93% of buried residues and 82-84% for all residues, and an RMSD of $\sim 1.7^\circ$ for side chain heavy atoms. The differences in prediction accuracy between the iterative and stochastic protocols are 1-2% for χ_1 and χ_1+2 predictions and about 0.1° in RMSD.

Software availability: The programs (Scomp-S and Scomp-I) are available for Unix/Linux platforms and also can be accessed through the web at <http://sgedg.weizmann.ac.il/scomp.html>. They can be used to model all, or a defined set of side chains in a protein as well as any number of user-generated changes (mutations). The programs can also employ a template PDB to place side chain conformations at conserved positions and model the remaining ones.

References:

1. Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J. Protein folding: the endgame. *Annu Rev Biochem* 1997;66:549-579.
2. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* 2001;311:421-430.
3. Mendes J, Baptista AM, Carrondo MA, Soares CM. Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants. *J Comput Aid Mol Des* 2001;15:721-740.
4. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol* 2002;320:597-608.
5. Mendes J, Nagarajaram HA, Soares CM, Blundell TL, Carrondo MA. Incorporating knowledge-based biases into an energy-based side-chain modeling method: application to comparative modeling of protein structure. *Biopolymers* 2001;59:72-86.
6. Liang S, Grishin NV. Side-chain modeling with an optimized scoring function. *Protein Sci* 2002;11:322-331.
7. Bower MJ, Cohen FE, Dunbrack RL. Prediction of protein side-chain rotamers from a backbone-dependent

rotamer library: a new homology modeling tool. *J Mol Biol* 1997;267:1268-1282.

8. Eyal E, Najmanovich R, McConkey BJ, Edelman M, Sobolev V. Importance of solvent accessibility and contact surfaces in modeling side chain conformations in proteins. *J Comp Chem* 2004;25:712-724.

9. von Freyberg B, Richmond TJ, Braun W. Surface area included in energy refinement of proteins. A comparative study on atomic solvation parameters. *J Mol Biol* 1993;233:275-292.

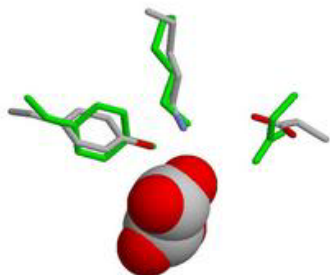
10. Zhou H, Zhou Y. Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. *Proteins* 2002;49:483-492.

Number: 17

Authors: **Gail J. Bartlett¹, James W. Torrance², Craig T. Porter, Jonathan A. Barker, Alex Gutteridge, Malcolm W. MacArthur, Janet M. Thornton**

¹Imperial College London, Centre for Bioinformatics, g.bartlett@imperial.ac.uk ²European Bioinformatics Institute

Title: **Generation of 3D Templates for Enzymes in the Catalytic Site Atlas: Analysis of Catalytic Residue Geometry and Utility of Automatically-generated templates for Function Prediction from 3D Structure**



Short abstract: 3D active site templates were automatically generated from enzyme structures in a database of catalytic sites. Templates were used to analyse catalytic residue geometry within homologous enzyme families, and also to test a template-based active site recognition method. Template effectiveness was measured using RMSD and statistical significance scores.

Full abstract: Pattern-matching and similarity-based methods have previously been used to successfully identify functional sites in protein structures. One of the disadvantages of such an approach is that it usually requires manual generation of templates, which can be labour-intensive and requires frequent reference to the literature to identify functionally significant residues. The Catalytic Site Atlas¹ is a database of enzymes with defined catalytic residues and their homologues and equivalent residues in the Protein Data Bank. This database was used to automatically create a set of 3D templates, based on atoms selected from the catalytic residues defining enzyme function. Two types of template were constructed: firstly, templates containing only C-alpha and C-beta atoms, and secondly, templates containing sets of three functional atoms per residue. Active site residue geometry within enzyme families contained within the Catalytic Site Atlas was analysed by superposing templates within the same family and examining the distribution of RMSDs between relatives. No correlation between RMSD of template atoms and percentage pairwise sequence identity of relatives was found, because RMSD of template atoms was strongly affected by other factors such as the presence or absence of ligand, type of ligand (substrate, transition state analogue, inhibitor), and changes to function, as well as by evolutionary divergence. The majority of RMSD values for superpositions between family members were below 1 Å, although some families did show wide variation in catalytic residue geometry, making these difficult to use for prediction.

The set of templates constructed in this manner made it possible, for the first time, to test the effectiveness of a template-based active site recognition method using a substantial set of templates. The Jess template-matching method² was used to query all the templates against a non-redundant dataset of the PDB. Hits were scored using both RMSD and statistical significance. The effectiveness of the templates was measured by comparing RMSDs between relatives with RMSDs from random hits in the PDB. In most families, templates distinguished related enzymes well. On average, more than 75% of relatives had RMSDs that were better than that of any random match.

Template-based methods are most useful in cases where methods based on sequence or overall structure would fail to detect a similarity between enzymes. The classic case of catalytic site similarity in the absence of overall structural similarity, between subtilisin and the serine proteases, is well known. However, there has been little study of how frequently such cases occur. The library of catalytic site templates has allowed the compilation of a set of cases where catalytic sites are similar despite differences in overall sequence and structure, as a result of evolutionary convergence or divergence.

References

1. Porter CT, Bartlett GJ, and Thornton JM (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* 32:D129-D133.

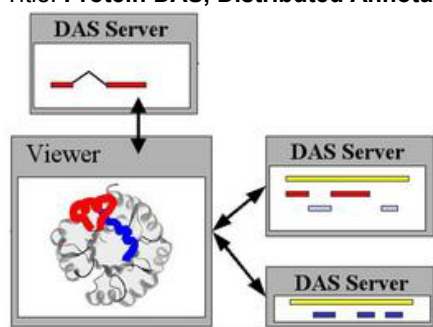
2. Barker JA and Thornton JM (2003). An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* 19:1644-1649.

Number: 19

Authors: **Andreas Prlic¹**, **Thomas A. Down**, **Tim JP. Hubbard**

¹The Wellcome Trust Sanger Institute, ap3@sanger.ac.uk

Title: **Protein DAS, Distributed Annotation System for Proteins**



Short abstract: DAS, the distributed annotation system, provides a communication protocol used to exchange annotations between decentralized data sources. We give an overview of extensions to the original DAS protocol in order to provide a client which displays DNA data like SNPs, or Intron and Exon borders, projected onto protein sequences and structures.

Full abstract: [DAS](#), the distributed annotation system, provides a communication protocol used to exchange biological annotations. It is based on the idea that annotations should not be provided by single centralized databases, but rather be spread over multiple sites. Data distribution, performed by DAS servers, is separated from the visualization, which is done by DAS clients. The currently most frequently used DAS client is the ENSEMBL ContigView e.g. [here](#)). The communication between the servers and clients is specified by the DAS-XML standards.

The original DAS protocol is designed to serve annotations of genomes. Here, it is extended to be able to deal with protein structures. An important problem when working with protein structures is, in which "coordinate system" annotations are being served. In contrast to DNA, or protein sequences, where the numbering is according to the sequence position, for protein structures the situation is more complicated. The PDB-residue numbering, which is one of the standards used by the community, is not guaranteed to be always linear or sequential. Annotations can be served in the coordinate system of a DNA or protein sequence, as well as in PDB-residue coordinates. Mapping services are required which define how one coordinate system can be mapped onto another.

The improvements in the DAS protocol allow new clients to be implemented. We are working on a program which can display DNA data like SNPs, or Intron and Exon borders, projected onto protein sequences and 3D protein structures. Other clients can be imagined which are based on new DAS services for multiple alignments of chromosomes, sequences and structures.

In this presentation poster we give an overview of the extensions to the DAS protocol and present the prototype of a client which can project features of the DNA to protein sequences and 3D structures.

Number: 20

Authors: **Ingolf E. Sommer¹**, **Joerg Rahnenfuehrer²**, **Francisco S. Domingues**, **Ulrik de Lichtenberg**, **Thomas Lengauer**

Max-Planck-Institute for Informatics, ¹sommer@mpi-sb.mpg.de ²rahnenfj@mpi-sb.mpg.de

Title: **Predicting Protein Structure Classes from Function Predictions**

First, a representative set of protein structures has to be constructed from the total 52000 deposited chain entries. To take account of drastic conformational changes observed in proteins during a functional screening, it is necessary to cover the largest possible structural variations of each protein. On the other hand, we cluster representative structures to reduce structural redundancy. Accordingly, we developed a protocol that allows us to extract representative sets of protein structures. The root of our selection is the accession number of proteins. Using additional data from our CUPS database (see figure) such as the resolution, R-factor, completeness, number of mutations, we identify a leading structure to which each of the rest RMSD lower than 0.5, form the α structures is compared. Structures having a C first cluster. The procedure is repeated for each remaining structure of the same protein.

Second, a selected structure needs a complete set of atomic coordinates. This requirement concerns integrity of structural coordinates and completeness of data found in the deposited entries. Molecular structures are atoms presented for α sometimes found in the following situations: only C amino-acid residues, or non-natural amino acids unmarked, or residues with a single atom missing. To prepare structures ready for screening using a molecular force field, we have developed a protocol that attempts to fix these problematic cases. We have also built coordinates of missing side-chain atoms in disordered area. In the current version, our program have repaired up to 94% of detected coordinate errors. We chose to work with the macromolecular Crystallography Information Files (mmCIF). To evaluate inconsistent items in all mmCIF files, we parsed the whole PDB and found an error rate of 19%. Weekly updates of modified mmCIF entries will be incorporated into CUPS, a MySQL relational database of cured protein structures.

Number: 23

Authors: Samantha L. Kaye¹, Mark SP Sansom², Philip C. Biggin

¹University of Oxford, samantha@biop.ox.ac.uk ²mark@biop.ox.ac.uk

Title: Simulation and Modelling studies of the NMDA Receptor



Short abstract: Ionotropic glutamate receptors (iGluR) are ligand-gated ion channels which mediate excitatory synaptic transmission. Crystal structures of the ligand binding domain of two types of iGluR have been solved. This poster describes the results of 30ns molecular dynamics simulations of one receptor (NR1), in complex with agonists and an antagonist.

Full abstract: Ionotropic glutamate receptors (iGluRs) are tetrameric ligand-gated ion channels which mediate excitatory synaptic transmission. Defined by their sensitivity to selective agonists, they are split into three subgroups, AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid), kainate and NMDA (N-methyl-D-aspartate).

The NMDA receptor is made up of two NR1 subunits, for which glycine is the ligand and full agonist, and two NR2 glutamate-binding subunits for which glutamate is the agonist. NR2 exists as one of four types known as A, B, C and D.

The crystal structures of the ligand binding domain of NR1 co-crystallised with 2 full agonists (glycine and D-serine), a partial agonist (D-cycloserine) and an antagonist (5,7-dichlorokynurenic acid) have been solved recently (Furukawa & Gouaux 2003). Using these structures as starting coordinates we have performed molecular dynamics simulations. Two additional models based on the crystal structures were also generated; a closed-apo, created by removing the glycine agonist from the crystal structure, and an open-apo, generated by removing the DCKA antagonist from the crystal.

The ligand binding domain of iGluRs is formed by two non-contiguous polypeptide chains (S1 and S2) which form a clamshell-like structure linked by a hinge region. The ligand binds in the pocket between the two sub-domains (domains 1 and domain 2). Previously, we have shown how different ligands induce different conformational dynamics within the AMPA receptor, GluR2. It is of interest to establish whether these observations reflect those of this fold type rather than being sequence specific.

Overall drift of the C α over time was calculated as a measure of the stability of the simulation. In each case there was an initial rise for 2.5 ns at which stage the RMSDs plateau at ~3 Å. Analysis of the results showed that D-serine exhibits lower overall C α RMSD than the glycine agonist, consistent with the suggestion from contacts in the crystal structure that D-serine binds more tightly to NR1. This is reflected in the inter-domain distance throughout the simulation. Further analysis was performed to investigate RMSD of domains relative to each other. This showed that domain 1 is more mobile than domain 2, supporting previous findings.

The ligand-protein interactions throughout the simulation were examined and the results compared. Of particular interest is the apparent opening of the ligand binding domain for the glycine bound structure toward the end of the 30 ns.

These simulation results will be related to the possible mechanisms of activation of NMDA channels.

Acknowledgements

Special thanks to the Medical Research Council (MRC) and to Oxford Supercomputer Centre.

References

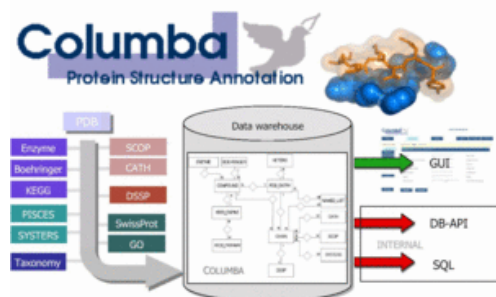
Furukawa, H. and Gouaux, E. (2003) Mechanisms of activation, inhibition and specificity: crystal structures of the NMDA receptor NR1 ligand-binding core. EMBO J., 22(12), pp. 2873-2885.

Number: 25

Authors: **Silke Trissl**¹, **Kristian Rother**², **Ulf Leser**

¹Institute of Informatics, Humboldt-Universität zu Berlin, Berlin, Germany, trissl@informatik.hu-berlin.de ²Institute of Biochemistry, Charité, Berlin, Germany, kristian.rother@charite.de

Title: **COLUMBA - A Database of Annotated Protein Structures**



Short abstract: We present **COLUMBA**, a database of information on protein structures that integrates data from twelve different biological databases, including the PDB, ENZYME, KEGG, SCOP, CATH, DSSP, and SwissProt. COLUMBA allows for the quick computation of sets of protein structures that share interesting properties according to the different data sources.

Full abstract: The number of protein structures deposited in the Protein Data Bank, PDB (Berman et al. 2000) is increasing rapidly. This allows researchers in life science to study complex relationships between macromolecular structures and their properties, such as biological function, folding classification, or secondary structure. To undertake those studies, not only the three dimensional (3D) structures have to be known, but also the folding classification and several other properties of a protein. Gathering such information from web resources by following hyperlinks is a tedious and time-consuming task.

We have created **COLUMBA**(Rother et al. 2004), a database of information on protein structures, that physically integrates information from twelve different data sources into a single relational data warehouse. We enrich the protein structures from the **PDB** with

- structure classifications from **SCOP** and **CATH**,
- computed secondary structures from the **DSSP** program,
- functional annotation from **ENZYME** and **GO**,
- participation in metabolic pathways from **KEGG** and the **Boehringer** map,
- taxonomic information from the **NCBI Taxonomy**,
- and further information from **SwissProt**.
- In addition to that, each chain is assigned to a cluster of similar sequences by **PISCES**.

Web interface

We have created a user friendly web interface, which is available at <http://www.columba-db.de>.

The web interface allows a full text search as well as data source specific queries. The full text search queries all available textual fields, e.g. the PDB header, folds from SCOP and CATH, enzyme, organism, and pathway names, and Gene Ontology terms. This search can be combined with more specific queries on the different data sources, for which own query forms are provided.

Our web interface uses a "query refinement" paradigm, which means that a result set obtained by conditions for one data source can be further restricted by conditions on a second to return the set of PDB entries with the desired properties. For this interactive and exploratory usage the user is guided by a header, which states the number of returned protein structures for each query step. The Results page shows the returned set of protein structures, including basic information on each entry. The user can see the full scope of **COLUMBA** for a single entry where all the annotated information to a PDB entry is shown. Apart from the information from the PDB itself, this includes the enzyme and pathway names for each compound as well as SCOP and CATH classification for each chain.

Applications

Apart from showing all possible information to one entry, **COLUMBA** allows to retrieve sets of protein structures, which share user defined properties. Such sets are essential for example for the computation of stereometric characteristics of amino acids within a certain fold, interaction of proteins in a molecular pathway, or for homology modelling. It is fairly simple to answer the following sample questions:

- Which structures contain chains with a TIM-barrel fold and have a resolution better than 2.0 Å.
- Which proteins in the citrate cycle do have a resolved structure?

The first query is answered by entering the phrase 'TIM barrel' as fold in the Protein Fold form of SCOP and '2.0' as restriction condition for the resolution in the PDB Structure form. This will result in the desired set of currently 416 protein structures for SCOP. The same query can be answered by using the Protein Fold form of CATH, which results in a set of 370 structures. It is even possible to restrict both CATH and SCOP to having a tim barrel fold. This yields 355 structures with a resolution better than 2.0 Å.

The second query can be answered by using the Metabolism form of **COLUMBA**. The option ?path coverage? not only shows the enzymes participating in the selected pathway, but also the number of structures known for each enzyme. For that particular pathway twelve out of 23 participating enzymes have at least one structure, which results in a coverage of 52%.

References

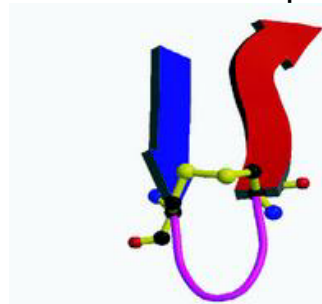
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Research*.28: 235-242.
 Rother, K., Müller, H., Trissl, S., Koch, I., Steinke, T., Preissner, R., Frimmel, C., Leser, U. 2004. COLUMBA: Multidimensional Data Integration of Protein Annotations. E. Rahm(Ed.): DILS 2004, LNBI 2994, 156-171.

Number: 26

Authors: **Merridee A. Wouters¹, Ken K. Lau, Philip J. Hogg**

¹Victor Chang Cardiac Research Institute, m.wouters@victorchang.unsw.edu.au

Title: **Cross-Strand Disulphides in Cell Entry Proteins: Redox Switches in Protein Nanomachines**



Short abstract: Cross-strand disulphides (CSDs) link adjacent beta-strands. A search for CSDs in a representative set of the PDB showed their dihedral strain energies are higher than disulphides in general. Conspicuously, CSDs are over-represented in molecules involved in cell entry and there is evidence suggesting their involvement in cell entry events.

Full abstract: Disulphides are generally viewed as structurally stabilizing elements in proteins. However, it is well known that some disulphides are redox active and capable of being reduced under physiological conditions. The enzymatic role of redox-active disulphides has been well characterized in thiol-disulphide reductases. It is less well known that similar disulphides are present in other, non-enzymatic, protein structures. Disulphide redox potentials measured in thiol-disulphide oxidoreductases range from -120mV to -270mV [1-4]. For disulphides serving structural purposes, the redox potential can be as low as -470mV [5]. Emerging evidence suggests that redox-active disulphides may act as switches of conformational change in proteins.

One such disulphide is the cross-strand disulphide (CSD). Cross-strand disulphides (CSDs) are unusual bonds that link adjacent strands in the same beta-sheet. Structurally, their peculiarity relates to the high potential energy

stored in these bonds, both as torsional energy in the highly strained disulphide linkage and as deformation energy stored in the sheet itself. Evidence that these disulphides may be redox-active include use of a CSD in the thiol-disulphide reductase DsbD [6]; demonstration that reduction of one of these disulphides in CD4 occurs during the entry of HIV gp120 into the cell [7]; and engineering of a redox-sensitive variant of green fluorescent protein by the introduction of a CSD into the structure [8,9].

Structural data mining is a powerful tool for uncovering common sub-structures in proteins. Initial characterization of the CSD was done as part of a structural bioinformatics survey of highly correlated pairs of cross-strand residues in beta-sheet [10]. The novelty of these structures was emphasized by the previous assertion that such structures would not exist because of the incompatible conformational constraints of beta-sheets and disulphides [11]. CSDs are always found in the 'non-H-bonded' site of antiparallel beta-sheet, an arrangement where the backbone hydrogen-bonding groups face away from each other. This site differs from the alternate 'H-bonded site' by having a smaller separation of the adjacent C_{alpha}s (4.5 Å compared to 5.5 Å) and not being directly constrained by two hydrogen bonds that the link the backbones of the cross-strand residue pair together. The disulphide linkage forms across the two strands roughly perpendicular to the strand direction and parallel to the hydrogen bonds between the two strands. Both half-cystines in the disulphide assume a highly stressed gauche + chi₁ conformation where the C_{alpha} and C_{beta} separations are roughly equal at 4 Å. The deformation of the sheet is mainly apparent as a pucker caused by the strands being drawn closer together and tilting towards each other to accommodate the disulphide linkage. Typically, C_{alpha} atoms of non-H-bonded residue pairs are separated by a distance of 4.5 Å, whereas CSD C_{alpha}s are separated by roughly 4.0 Å. In addition, the strands usually shear in the direction of their C-terminus, as well as twisting in the usual direction.

Studies of disulphide-bonds in model proteins have shown that those bonds with high potential energy are more easily cleaved than disulphide-bonds with lower stored energy [12-15]. The strain of a disulphide-bond can be estimated from the five dihedral angles [12,13]. The calculated strain energies only consider the dihedral angles and do not include other factors such as bond lengths, bond angles, and van der Waals contacts in calculating energy. Nevertheless, the findings of Pjura et al [16] indicate that such calculations can give useful semi-quantitative insights into the amount of strain in a disulphide-bond. We calculated dihedral strain energies for all disulphides in a representative set of the PDB and compared them to the energies of CSDs. CSDs have an average torsional energy of 18.8 ± 6.2 kJ.mol⁻¹ (208 disulphides in 171 proteins), which is approximately 3.5 kJ.mol⁻¹ higher than average. Significantly, they represent 8% of disulphide bonds that have torsional energies higher than 12.5 kJ.mol⁻¹ and are basically unrepresented at lower energies.

CSDs are relatively rare in protein structures but are conspicuously over-represented in molecules involved in cell entry. The finding that entry of botulinum neurotoxin and HIV-1 into mammalian cells involves cleavage of CSDs suggests that the activity of other cell entry proteins may likewise involve cleavage of these bonds. 29 CSDs are found in 14 proteins involved in cell entry, which equates to an average of 2.1 CSDs per protein. In contrast, the frequency of a CSD in any disulphide-containing protein is 0.13. We present evidence suggesting the involvement of these unusual disulphides in cell entry events [17].

Correspondence to m.wouters@victorchang.unsw.edu.au

References:

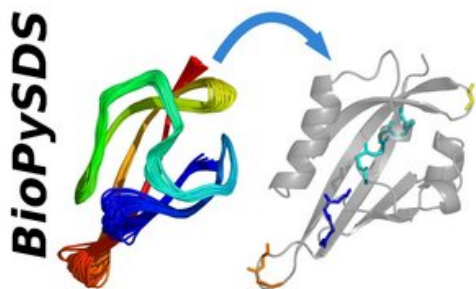
- [1] Wunderlich M, Glockshuber R. *Protein Sci.* 1993 May;2(5):717-26.
- [2] Huber-Wunderlich M, Glockshuber R. *Fold Des.* 1998;3(3):161-71.
- [3] Lin TY, Kim PS. *Biochemistry.* 1989 Jun 13;28(12):5282-7.
- [4] Krause G, Holmgren A. *J Biol Chem.* 1991 Mar 5;266(7):4056-66.
- [5] Gilbert HF. *Adv Enzymol Relat Areas Mol Biol.* 1990;63:69-172.
- [6] Goulding CW, Sawaya MR, Parseghian A, Lim V, Eisenberg D, Missiakas D. *Biochemistry.* 2002 Jun 4;41(22):6920-7.
- [7] Matthias LJ, Yam PTW, Jiang X-M, Vandegraaff N, Li P, Pountourios P, Donoghue N, Hogg PJ. *Nature Immunol.* 2002;3:727-732.
- [8] Ostergaard H, Henriksen A, Hansen FG, Winther JR. *EMBO J.* 2001 Nov 1;20(21):5853-62.
- [9] Hanson GT, Aggeler R, Oglesbe D, Canon M, Capaldi RA, Tsien RY, Remington SJ. *J Biol Chem.* 2004 Jan 13
- [10] Wouters MA, Curmi PM. *Proteins.* 1995 Jun;22(2):119-31.
- [11] Richardson JS. *Adv Protein Chem.* 1981;34:167-339.
- [12] Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta SJr, Weiner P. *J. Am. Chem. Soc.* 1984;106:765-784.
- [13] Katz BA, Kossiakoff A. *J. Biol. Chem.* 1986;261:15480-15485.
- [14] Wells JA, Power DB. *J. Biol. Chem.* 1986;261:6564-6570.
- [15] Wetzel R. *Trends Biochem. Sci.* 1987;12:478-482.
- [16] Pjura PE, Matsumura M, Wozniak JA, Matthews BW. *Biochemistry* 1990;29:2592-2598.
- [17] Wouters MA, Lau KK, Hogg PJ. *Bioessays.* 2004 Jan;26(1):73-9.

Number: 27

Authors: **Alessandro Pandini**¹, **Giancarlo Mauri**², **Laura Bonati**

¹DISAT - Universita' degli Studi di Milano-Bicocca, alessandro.pandini@unimib.it ²DISCO - Universita' degli Studi di Milano-Bicocca, giancarlo.mauri@unimib.it

Title: **BioPySDS: an Object-Oriented Interface to Manage Protein Dynamics in an Evolutionary Framework**



Short abstract: A new approach to computational investigation of protein function is presented: structural, dynamical and evolutionary informations are combined and compared to gain deeper insight in the function of protein superfamilies. **BioPySDS**, a new object oriented tool is presented: design, implementation and application are detailed.

Full abstract: Protein functions could be addressed with higher degree of accuracy by comparative analysis of the dynamical, structural and evolutionary properties of their domains. Comprehension of the dynamical features within an evolutionary framework could highlight how natural selection acted at macromolecular level to conserve, specialize and take advantage of key conformational changes for biological mechanisms.

Whereas protein structures and sequences have been generated by experimental and theoretical methods and collected in worldwide available databases, fewer information have been derived on protein dynamics. Thanks to more and more powerful computers and more accurate force fields and scoring functions, molecular simulations are emerging as effective tools to investigate dynamical properties also for relatively large systems as proteins. Among these methods, Molecular Dynamics coupled with Covariance Analysis has been identified as an excellent tool to extract essential motions with putative functional interpretation. This approach partially fills the gap of information about dynamics of macromolecules. Anyway the recent availability of data do not cover the lack of tools to manage dynamics information on protein structures with an evolutionary approach.

In this work the design and implementation of a core set of modules is presented that allows a deeper comprehension of protein function through comparative analysis of structural, dynamical and evolutionary properties.

Dealing with different levels of information, represented by fairly different sources and data structures, required the employment of a flexible and high-level modeling approach: both design and implementation were done within an object oriented paradigm. As a first step in the software development process, an UML description of the problem was generated: a set of classes that suitably model the different levels of information and their relationships was devised and detailed.

The UML model was then translated to a *python* implementation. The choice of *python* was motivated by its flexibility, elegance and native object oriented support. Availability of a large number of built-in general purpose classes was an additional benefit that supported the choice. In this direction, both in the design and implementation processes *Biopython* (<http://www.biopython.org>) modules were employed as references for code styling and file parsing modes. Sequence classes were directly derived from *Biopython* whereas new extensions were designed and implemented for Trajectory, Frame, Property and Fragment representations. Trajectory and Frame classes model the traditional Molecular Dynamics objects with the ability to manipulate initial structure and simulation frames, and to run basic geometrical analyses directly, whereas Property and Fragment are aimed to manage internal and external processed data. Property offers different general containers for properties of atoms, residues and time. Fragment represents the concept of variable length substructures and allows direct access to parts of the protein structure that fulfill specific Property requirements. To offer smooth access to sequence representation at each stage of data processing, the Sequence counterparts of the initial structure in Trajectory and the derived Fragment objects are fully embedded in the implementation. Additionally a general parser for MD trajectory and property files was developed and a specific parser for GROMACS¹ files was coded. The *Biopython* PDBParser and SMCRA (Structure-Model-Chain-Residue-Atom) model was employed to describe structures.² Interrogations to Protein Databank were implemented in a PDBConnector class with a flexible retrieving system to generate both Structure and Fragment objects. The developed set of modules was named **BioPySDS BioPython Structure-Dynamics-Sequence interface**) after the pythonic implementation and the accessible information levels.

BioPySDS was firstly employed to investigate the dynamical features of a superfamily of receptors with signal transmission functionality, the PAS protein collection. These are multi-domain proteins characterized by low sequence similarity but with a highly conserved structural unit encompassing a 100 aminoacids domain. A set of

40 ns MD simulations were previously generated by GROMACS for the four representative structures of PAS domains and a suitable index of structural mobility was devised in the RMSF on C-alpha positions derived for a reduced Essential Subspace of 6 dimensions.

BioPySDS was employed to identify high motion substructures and derive the corresponding Sequence and Fragment objects. Representation and analysis of RMSF values on a Fragment base allowed identification of structural patterns of motion for each PAS domain. Comparison of them across the PAS superfamily extended the concept of structural motion patterns to the superfamily level. From this analysis, three main regions with aligned motions were identified: two supportive N-term and C-term part and a central functional part which exploits also a specific secondary structure arrangement in agreement with the latest hypotheses on PAS receptive mechanism.

Next step on **BioPySDS** development will consist in exploiting information from dynamics to improve low similarity sequence searches and alignment. In this direction motion patterns will be used to weight PSSMs for PSI-BLAST³ searches and to derive Substitution Matrices for sequence alignments.

Overall **BioPySDS** is a first effort to offer well-designed, general-purpose solutions to analyze domain functions in protein superfamilies through different levels of information. Its main novelty consist in the idea of using information from the analysis of protein dynamics in an evolutionary framework.

¹ Lindahl, E., Hess, B. and van der Spoel, D., "GROMACS 3.0: A package for molecular simulation and trajectory analysis" *J. Mol. Mod.* (2001) **7**:306-317.

² Hamelryck, T. and Manderick, B., "PDB file parser and structure class implemented in Python" *Bioinformatics* (2003) **19**:2308-2310.

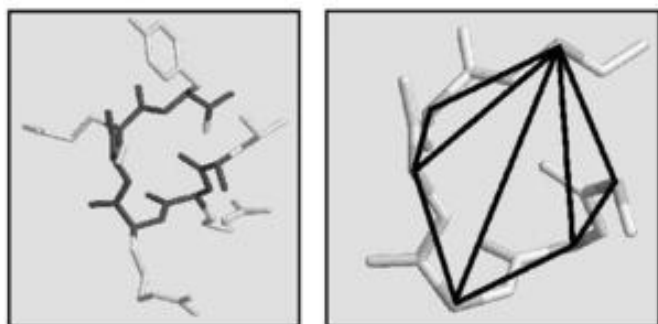
³ Altschul, S.F., Madden, T.L., Schaeffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* (1997) **25**:3389-3402.

Number: 29

Authors: Chen Yanover¹, Ori Shachar², Yair Weiss

The Hebrew University of Jerusalem, ¹cheny@cs.huji.ac.il ²orish@cs.huji.ac.il

Title: Inference in Graphical Models for Side-chain Prediction



Short abstract: Inference in graphical models is a widely studied problem in computer science. We apply various inference algorithms to the side chain prediction problem, both for finding the lowest energy configurations and characterizing the variability in configurations, and show excellent results. An extension to protein peptide binding prediction gives promising results.

Full abstract: Many structural biology problems can be represented as a set of discrete variables interacting with each other locally. Side-chain prediction, that is predicting side-chain conformation given the backbone structure, is such a problem – the use of rotamer library makes the conformational space discrete and pairwise energy terms keep interactions local. Side-chain prediction is a central problem in protein-folding and molecular design. It arises both in ab-initio protein-folding and in homology modeling schemes.

Graphical models are data structures that represent the qualitative nature of a probability distribution by means of a graph. Inference in a graphical model, namely answering probabilistic queries, is a widely studied problem in computer science and statistics. A large number of inference algorithms have been developed and successfully applied to many problems. In this work, we ask whether these algorithms can be used to improve the state of the art in side-chain prediction.

Using a database of hundreds of proteins we empirically assessed the performance of exact and approximate inference algorithms. For the task of finding the lowest energy (that is the most probable) configuration standard exact inference algorithms combined with dead end elimination (DEE) obtain the global minimum in few seconds for the vast majority of proteins. In terms of characterizing the distribution of low energy configurations, exact inference is intractable for most proteins but approximate inference algorithms give excellent performance. For

tasks that involve characterizing the variability in configurations (e.g. calculating the top M configurations, calculating the variability of a single residue and estimating the conformational entropy and free energy) the graphical model approach appears to outperform many existing algorithms.

Finally, we try using the side-chain prediction utility for predicting protein-peptide complex affinity, assuming that protein and peptide backbones are rigid. Preliminary results show that approximate free energy, calculated using the graphical model representation, can predict binding peptides for various domains.

Number: 30

Authors: **Fabrizio Ferre**¹, **Gabriele Ausiello**², **Manuela Helmer-Citterich**³, **Andreas Zanzoni**

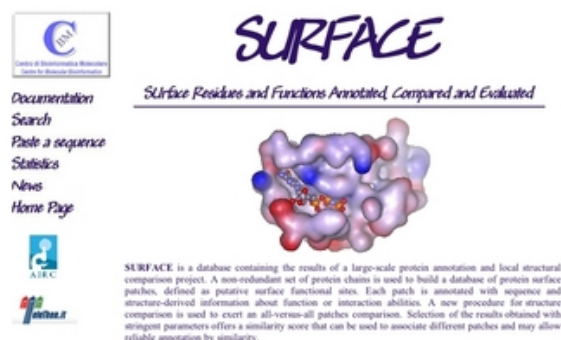
¹Centre of Molecular Bioinformatics, University of Rome Tor Vergata and Boston College, Boston MA,

fabrizio@cbm.bio.uniroma2.it ²Centre of Molecular Bioinformatics, University of Rome Tor Vergata,

gabriele@cbm.bio.uniroma2.it ³Centre of Molecular Bioinformatics, University of Rome Tor Vergata,

citterich@uniroma2.it

Title: **Large Scale Surface Comparison for the Identification of Functional Similarities in Unrelated Proteins**



Short abstract: A systematic local comparison of protein surfaces is performed aiming at functional annotation based upon stringent shape and residue similarity criteria. Protein surface patches and the results of the comparison are stored in the [SURFACE](#) database. Significant similarities are found between proteins sharing low sequence or structural homology (structural genomics).

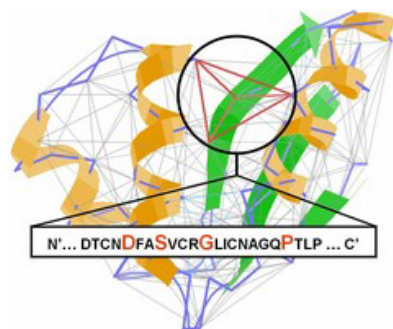
Full abstract: We developed a systematic large-scale approach to identifying protein surface regions sharing shape and residue similarity. We used a new fast structural comparison algorithm (LSC: Local Surface Comparison) to exhaustively analyze a set of functionally annotated protein patches ([Laskowski, 1995](#)) with a larger collection of protein cavities. From a dataset of about 10.000 protein surface patches extracted from a non redundant list of PDB proteins (about 2000 chains), we collected a grand total of 65910 matches among patch pairs that were stored in the [SURFACE](#) ([Ferre et al., 2004](#)) database. The functional meaning of most of the matches could be confirmed by other established methods: the presence of the same PROSITE ([Hulo et al., 2004](#)) and ELM ([Puntervoll et al., 2004](#)) motifs in the sequence, the presence of the same ligand in the PDB ([Bourne et al., 2004](#)) structure, similar GO ([Harris et al., 2004](#)) terms, common SWISS-PROT ([Boeckmann et al., 2003](#)) keywords, sequence similarity, same SCOP ([Andreeva et al., 2004](#)) superfamily and E.C. ([Bairoch, 2000](#)) numbers. We noticed that the fraction of matches whose functional association can be confirmed by more methods sensibly decreases with the extension of the match ([Figure1a](#)). Interestingly, a considerable fraction of the more significant matches (zscore>7) occurs between protein patches belonging to different SCOP folds ([Figure1b](#)). So by considering only highly significant matches, we could characterize a number of proteins solved in structural genomics projects and annotated as being of unknown function. The strategy we developed is a powerful addition to the list of available methods for the functional annotation of structures of unknown function.

Number: 32

Authors: **Ruchir R. Shah**¹, **Luke Huan**², **Deepak Bandyopadhyay**, **Wei Wang**, **Alexander Tropsha**

University of North Carolina at Chapel Hill, ¹ruchir@email.unc.edu ²huan@cs.unc.edu

Title: **Structure Based Identification of Protein Family Signatures for Function Annotation**



Short abstract: We present a novel approach to identifying recurrent structure-sequence motifs common to particular protein families. The approach employs Delone tessellation to generate protein graphs and frequent subgraph mining to obtain the motifs. We demonstrate the utility of these motifs for highly accurate annotation of several protein families.

Full abstract: Protein sequences are known to evolve much faster than structures. Pairs of distantly homologous proteins with very low sequence similarity may still share very similar folds and functions. One example of well-preserved structural motifs is given by specific three-dimensional orientations of amino acid residues in the active site of a protein, which are responsible for protein's biochemical functions. However, most of the time active site residues are hard if not impossible to infer from the primary sequence since the active site residues are frequently quite distant from one another. Developing automated approaches to identifying protein family signatures can be very useful in protein family classification and function annotation.

We present a novel automated approach to identifying recurrent structure-sequence motifs common to particular protein families. We use a labeled graph to represent the structure of a protein where each amino acid residue is represented by a $C\alpha$ node. Two independent approaches were used to generate edges. In the first approach two nodes are connected by an edge if the two residues are either connected by a peptide bond or the distance between them was below a certain threshold in the absence of a peptide bond. Whereas in the second approach the edges result naturally from applying computational geometry approach, i.e., Delone tessellation to $C\alpha$ nodes in 3D space. We then employ a fast and efficient frequent subgraph mining algorithm developed in our group to extract common packing patterns (i.e., subgraphs) from a graph family corresponding to a protein family in the SCOP database. Finally, we examine these common substructural packing patterns for the presence of conserved sequence patterns which thereby represent family specific structure-sequence signatures similar to Prosite motifs. These signatures are used to query protein sequence databases and the hits (i.e., sequences that contain the signatures) are predicted to belong to the same structural/functional class as the protein family from which the signatures were derived.

The Serine Protease and Nuclear Receptor Binding protein families were used as test cases. A diverse subset of structures belonging to a given family was selected from the SCOP database such that the pair-wise sequence identity between any two proteins in the set was below a certain threshold. Subgraph mining was performed to extract recurring structural motifs specific to the given family. Motifs were investigated for the location of residues in the three dimensional structures by superposing structures in SwissPDBviewer. Furthermore, amino acid residues in the motifs were analyzed for their order in the primary sequence and the lengths of the loops separating residues in the primary sequence were calculated.

Function-specific sequence signatures were derived and used to search the Swissprot sequence database. A protein sequence was called a hit if it contained the query functional signature and hits were analyzed for precision (True Positive/[True Positive + False Positive]) and recall (True Positive/[True Positive + False Negative]) values. Results for functional signatures derived from trypsin like serine protease are summarized below; for comparison, we have conducted similar searches using known Prosite motifs for the same family.

Motif	Precision	Recall
C-G-x{11}-A-A-H-C	100%	75%
D-S-G-G-P	93%	90%
Prosite		
PS00134*	95%	88%
PS00135**	98%	88%
*PS00134: [LIVM]-[ST]-A-[STAG]-H-C		
**PS00135:[DNSTAGC]-[GSTAPIMVQH]-x{2}-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH]		

In addition, we used two above motifs and two Prosite motifs simultaneously for sequence searches. Here, a protein was considered a hit if it contained at least one of these motifs. Searching with the pair of motifs discovered by our approach yielded 93% precision and 91% recall; whereas the pair of Prosite motifs gave 93% Precision and 95% Recall. It is interesting to note that, the query using our motifs identified five trypsin like serine proteases (Swissprot IDs: VSP2_TRIFL, VSP4_AGKAC, VSPA_LACMU, VSP1_AGKRH, VSP2_AGKRH) that were missed by the query using the Prosite motifs.

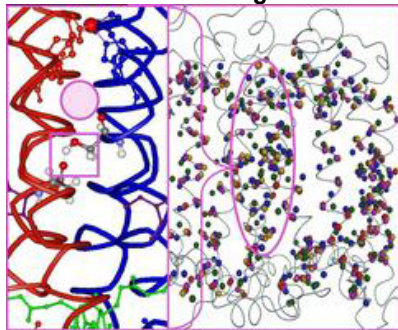
The results indicate that our structure based approach is able to annotate protein sequences with accuracy comparable to that of other sequence-based methods such as Prosite. However, our motifs appear more compact (see Table) and are derived from much smaller set of proteins with known structure. In some cases we were able to correctly predict the protein functions that Prosite was unable to annotate. This structure-based approach is currently being extended to all major protein families in the SCOP and similar databases. The results of this study shall provide a diverse set of family specific signatures that can be used for the genome-wide annotation of protein functional families.

Number: 33

Authors: Ilan Samish¹, Eran Goldberg², Oksana Kerner, Avigdor Scherz

¹Weizmann Institute of Science, ilan.samish@weizmann.ac.il ²eran.goldberg@weizmann.ac.il

Title: Centrality of Weak Interhelical H-bonds in Membrane Protein Functional Assembly and Conformational Gating



Short abstract: We show (see ISMB/ECCB2004 [paper 57](#)) that weak, membrane protein backbone-mediated interhelical H-bonds cluster in the conserved, buried protein core. The residue-propensity for these bonds correlates with the packing scale. *In-silico* and *in vivo* mutagenesis of such bonds suggests a mechanism for the conformational-gated electron-transfer in photosynthetic reaction-centers. The results provide insight for structure and function prediction.

Full abstract: Folding of helical membrane proteins is constrained by a balance of forces, namely the need to remain hydrophobic during helix insertion into the membrane, the need to include polar residues in support of electrostatic interhelical interactions, and the need to maintain conformational flexibility required for protein function. Backbone-mediated interhelical H-bonds may provide a tool to compensate between these contradicting needs.

Our analysis (see ISMB/ECCB2004 [short paper 57](#)) demonstrates that backbone-mediated interhelical hydrogen-bonds cluster in the conserved, central core of transmembrane helical proteins. Each residue's propensity to bear these interactions is in correlation with the residue's packing-value scale; giving biophysical meaning to this phenomenological scale. Residues participating in such an intersubunit, structurally conserved H-bond in reaction centers of photosystem II were combinatorially mutated and characterized *in silico* and *in vivo* suggesting that H-bond reversible association regulates protein-gated electron transfer. Similar motifs were found in key residues of Rhodopsin and of the K⁺ channel and may be involved in folding and conformational flexibility of other membrane proteins. Hence, these findings provide new parameters for structure and function prediction.

Our study integrates bioinformatic tools in a feedback loop – guiding the genetic manipulation strategy and explaining the resulting biophysical findings. Generalization of key findings in the model system is presented by searching for the phenomenon in sequence and structure space and vice versa. Finally, the recent accumulation of helical membrane protein structures enabled us to conduct statistical datamining of the phenomenon suggested by our model system in a non-redundant dataset of all helical membrane proteins. Cumulatively, our results suggest new roles for weak interhelical H-bonds in the membrane milieu as well as biophysical meaning to interactions previously regarded phenomenologically as 'packing'.

Bioinformatics methods applied in our study include sequence motif alignment and conservation, structure conservation (multiple structural alignment), in-house program for H-bond detection, rotamer-library based mutagenesis, and intra-protein cavity analysis.

Number: 34

Authors: Dmitry S. Kanibolotsky¹, Konstantin A. Odyets², Alexander I. Kornelyuk³

Title: **Homology Modeling and Molecular Dynamics Simulation Study of Cytokine-like C-terminal Module of Mammalian Tyrosyl-tRNA Synthetase**



Short abstract: Homology modeling of 3D structure and exploration of molecular dynamics of tyrosyl-tRNA synthetase C-module were performed. Our data revealed that 86-92 and 107-117 loops are most flexible parts of protein structure. The dynamics of cytokine motif revealed the increase of Arg8 and Arg12 solvent accessibilities.

Full abstract: The C-terminal non-catalytic module of mammalian tyrosyl-tRNA synthetase reveals high sequence homology with proinflammatory EMAP II cytokine and also displays several cytokine activities in vitro. On other hand, the C-module is a structure-specific tRNA-binding domain. Earlier the hypothesis was proposed that the conformation of C-module is flexible and diverse functions of protein could be realized in different conformations. With the aim to explore a conformational flexibility of this module we performed a comparative homology modeling of 3D structure and molecular dynamics (MD) simulation studies of bovine tyrosyl-tRNA synthetase C-module in solution. The model of the TyrRS C-module structure has been built with Swiss-Model server and Modeller v.4 program by homology modeling approach using known crystal structures of EMAP II cytokine and human C-TyrRS. The model represents a globular protein of 166 amino acids which is composed of the OB-fold domain linked to the appended A-subdomain. It was found that Pro6-Ser18 motif responsible for cytokine activity is located at the first beta-strand of the OB-fold. The MD simulation was carried out using Gromacs 3.1.4 program package. Protein was placed into a box with minimal distance to the walls of 1 nm with periodic-boundary conditions and the box was filled with water molecules. Two Na⁺ ions were added into system for charge neutralization and 2000 steps of the energy minimization was carried out using steepest-descent algorithm. The actual MD simulation of free unrestrained protein was conducted for 4.4 ns time range. These MD data were compared with Ca-atoms fluctuations derived from crystallographic B-factors of template structures and a good correlation was observed. Our data revealed that two loops 86-92 and 107-117 were most flexible parts of the C-module structure. These loops are proposed to be involved in the structure-specific tRNA binding by C-module. The dynamical behavior of cytokine Pro6-Ser18 motif was analyzed during 4.4 ns. It was revealed that Arg8 and Arg12 residues of this cytokine motif increased their solvent accessible surfaces during MD simulation. Our data suggest that Arg8 and Arg12 residues are most important for cytokine activity of TyrRS C-module and possibly play a key roles in the receptor binding.

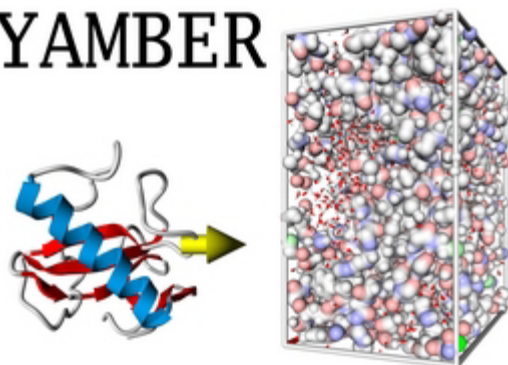
Number: 35

Authors: **Elmar Krieger**¹, **Tom Darden**², **Sander B. Nabuurs**, **Alexei V. Finkelstein**, **Gert Vriend**

¹University of Nijmegen, elmar@cmbi.kun.nl ²NIEHS

Title: **Making Optimal Use of Empirical energy Functions: Force Field Parameterization in crystal Space**

YAMBER



Short abstract: Increasingly accurate energy functions are required to improve protein models built by homology, for example during molecular dynamics refinement. By simulating protein crystals and adjusting the AMBER force field parameters to minimize the deviations from experiment, we arrive at the YAMBER force field, which is shown to improve homology models.

Full abstract: Today's energy functions are not able yet to distinguish reliably between correct and almost correct protein models. Improving these near-native models is currently a major bottle-neck in homology modeling or experimental structure determination at low resolution. Increasingly accurate energy functions are required to complete the 'last mile of the protein folding problem', for example during a molecular dynamics simulation.

We present a new approach to reach this goal. For 50 high resolution X-ray structures, the complete unit cell was reconstructed, including disordered water molecules, counter ions and hydrogen atoms. Simulations were then run at the pH at which the crystal was solved, while force field parameters were iteratively adjusted so that the damage done to the structures was minimal. Starting with initial parameters from the AMBER force field, the optimization procedure converged at a new force field called YAMBER (Yet Another Model Building and Energy Refinement force field), which is shown to do significantly less damage to X-ray structures, often move homology models in the right direction and occasionally make them look like experimental structures.

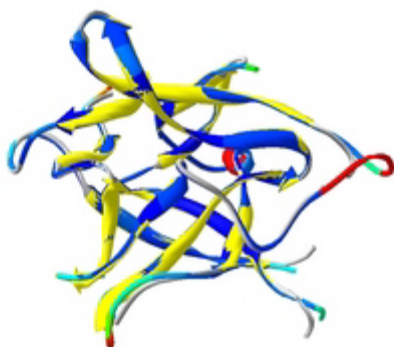
Application of YAMBER during the CASP5 structure prediction experiment yielded a model for target 176 that was ranked first among 150 submissions. Due to its compatibility with the well established AMBER format, YAMBER can be used by almost any molecular dynamics program. The parameters are freely available from www.yasara.org/yamber.

Number: 37

Authors: Antonis Koussounadis¹, David W Ritchie², Antonis Koussounadis, Chris J. Secombes

¹Department of Computing Science/ Zoology, University of Aberdeen, akoussou@csd.abdn.ac.uk ²Department of Computing Science, University of Aberdeen, dritchie@csd.abdn.ac.uk

Title: Modelling b-Trefoil Proteins Using an Object-Oriented Database



Superposition of modelled trout IL-1 on human

Short abstract: This presentation describes on-going work using an object-oriented database, P/FDM, to build 3D homology models of b-trefoil proteins, by exploiting existing structural data. The model-building procedure is based on the querying capabilities of P/FDM that allows data access, navigation, computation to be mixed easily in order to formulate structural queries.

Full abstract: This presentation describes on-going work using an object-oriented database, P/FDM, to build 3D homology models of b-trefoil proteins such as Interleukin-1 (IL-1), by exploiting existing structural data from known structures of the same fold. The model-building procedure is based on the powerful querying capabilities of P/FDM that allows data access, navigation, and computation to be mixed freely and easily in order to formulate complex structural queries. The database can be queried using a Java-based graphical user interface. Using the P/FDM b-trefoil database, the above approach allows to "cut-and-paste" fragments from existing b-trefoil structures to build the model. The structural part of the database contains objects for the description of proteins, as well as b-trefoil structure-specific entities. Data are organised according to a global b-trefoil structural alignment based on superposition of the 18 core residues of the fold, identified from previous structural studies. The database is currently populated with 56 trefoil domains and further structures are being added in order to make available a richer collection of "spare parts" for modelling. Future enhancements such as improved side chain placement and fragment clustering are also described.

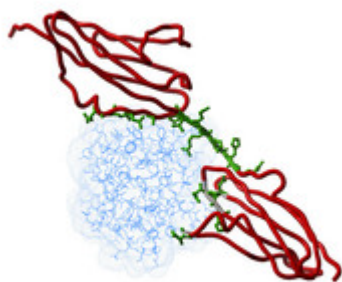
Preliminary protein models of fish IL-1 that adopt a b-trefoil fold, prepared according to our method, will be described and compared to their human homologues. The significance of the differences among various fish IL-1 will be discussed with a view to identify species-specific ways of receptor engagement and functionally important regions.

Number: 38

Authors: Juan Fernandez-Recio¹, Tom L. Blundell²

University of Cambridge, ¹juan@cryst.bioc.cam.ac.uk ²tom@cryst.bioc.cam.ac.uk

Title: Computer Simulations of Protein-Protein Interactions: Rigid-Body Docking and Flexible-Link Refinement Applied to Multi-Domain Signalling Complexes



Short abstract: The ICM-DISCO method for protein-protein docking was recently validated in a blind assessment (CAPRI). The method provides a description of the rigid-body association energy landscape and identifies protein-binding areas on protein surfaces. As a further step, a new flexible-link refinement protocol is applied here to multi-domain signalling complexes.

Full abstract: Accurate modeling of protein-protein or domain-domain interactions is one of the challenges of the post-genomic era. Once the entire genomes of many organisms (including human) have been sequenced, the goal now is to determine the structure of all gene products (i.e. proteins), with the immediate purpose of understanding how these proteins interact with each other for their biological functions. As experimental techniques are still unable to systematically produce accurate models for protein-protein complexes, the need of theoretical physical models is increasing. Among the emerging protein-protein docking methods, the ICM Docking and Interface Side-Chain Optimization (ICM-DISCO) was benchmarked in 24 unbound pairs [1] and presented one of the highest predictive rates in the blind CAPRI experiment (<http://capri.ebi.ac.uk>) [2]. In addition to the ability of making successful predictions of biological interest, the method provides an accurate description of the association energetics. Thus, analysis of the docking landscapes generated from these pseudo-Brownian rigid-body simulations indicates the existence of preferred protein-binding areas on protein surfaces [3]. These studies provide an interesting image of the binding energy landscape where desolvation and electrostatics are the driving forces for the association process. Recently, we have begun to explore the application of these methods to protein-protein complexes involved in signal transduction. A detailed energetic and structural knowledge of these complexes is needed in order to understand the complex network of signalling pathways, and also to design drugs for modulating interactions of therapeutical interest. In the case of the FGF:FGFR interaction, a family of more than 20 growth factors have varying affinities for the almost 20 splice forms of the five FGF receptors. We have thus devised an approach for exploration of the conformational space by a new developed protocol for 'linked-tethered' docking, in which domains D2 and D3 in FGFR are treated as rigid-bodies (e.g., fixed backbone and side-chain atoms), and the link between them is fully flexible. Loose restrains are imposed to keep one of the domains in the vicinity of the starting conformation during the simulations, while the other domain can move freely with the only conformational restrictions that the flexible link impose. Using this methodology, we are able to reproduce the affinity of the FGF:FGFR interaction for some of the members of the FGF family [4]. The implications for the specificity of the different members of the FGF/FGFR system will be discussed.

[1] Fernandez-Recio, J., Totrov, M., Abagyan, R. (2002) Soft protein-protein docking in internal coordinates. *Protein Sci.* 11, 280-291

[2] Fernandez-Recio, J., Totrov, M., Abagyan, R. (2003) ICM-DISCO Docking by Global Energy Optimization with Fully Flexible Side-Chains. *Proteins* 52, 113-117

[3] Fernandez-Recio, J., Totrov, M., Abagyan, R. (2004) Identification of Protein-Protein Interaction Sites from Docking Energy Landscapes *J.Mol.Biol.* 335, 843-865

[4] Harmer, N.J., Pellegrini, L., Chirgadze, D., Fernandez-Recio, J., Blundell, T.L. (2004) The crystal structure of Fibroblast Growth Factor (FGF) 19 reveals novel features of the FGF family and offers a structural basis for its unusual receptor affinity *Biochemistry* 43, 629-640

Number: 39

Authors: Hugh P. Shanahan¹, Sue Jones², Carles Ferrer, Janet M. Thornton

¹EBI, Hugh.Shanahan@physics.org ²University of Sussex, S.Jones@sussex.ac.uk

Title: Detecting DNA-binding Proteins using Structural Motifs and Electrostatics



Short abstract: We present a method for detecting DNA-binding proteins by a structural motif search and a positive electrostatic potential in that region. We demonstrate that this approach has a true positive rate that is comparable to more complicated structural methods of detecting DNA-binding proteins.

Full abstract: One of the challenges of Structural Genomics is to elucidate the function of proteins of known sequence and unknown function. In this poster we focus on methods for identifying the fraction of proteins that bind to DNA. This is a non-trivial task as it has been estimated that 6-7% of all eukaryotic proteins bind DNA¹. While there are a number of possible parameters that can be used to identify a DNA-binding protein, we combine searches for a set of structural motifs and a positive electrostatic potential on the surface of a putative DNA-binding protein. This approach is only relevant if a three dimensional structure is available. It has been observed that many known DNA-binding proteins have one of a small number of distinct structural motifs that play a key role in binding DNA². In particular, we focus on the Helix-Turn-Helix (HTH) motif, which one of the most common motifs used by proteins to bind DNA. The HTH motif is found in approximately one-third of DNA-binding protein structure families (16/54 families identified by Luscombe *et al.*³), and hence represents a significant proportion of proteins that bind DNA.

Likewise, it is well known that, given the distinct negative charge distribution between the backbone and bases of DNA, a DNA-binding protein has a complementary positive electrostatic potential in its binding region. This was used initially to identify the DNA-binding nature of the Tubby protein⁴. The calculation of an electrostatic potential and its use in the prediction of DNA-binding sites has previously been presented⁵. Each accessible atom is assigned a score which is proportional to the surface integral of the potential over a region projected from the accessible surface which is 7 Angstrom from each atom.

In order to search for a DNA-binding protein with a HTH motif, we apply three tests. We examine if a protein has a sequentially contiguous structural fragment which has a sufficiently small RMSD to a set of templates constructed from the observed DNA-binding proteins with a HTH motif. If such a fragment exists, we ask if it is sufficiently accessible. We also ask if the electrostatic potential in the region of the motif is sufficiently positive. This is carried out by computing the sum of electrostatic scores, described by Jones *et al.*⁵, over the accessible atoms that are part of the fragment.

We examine a non-redundant set of DNA-binding HTH protein structures and a non-redundant set of structures from the PDB. Using a simple linear predictor to divide the sets, we find that using these parameters are quite complementary and find a comparable true positive rate (sensitivity) to those obtained using more complicated structural methods of detecting DNA-binding proteins⁶. We find that the method has also a high accuracy and specificity.

Finally, we discuss the possibility of genome-wide scans and how it can be employed to complement HMM searches for Transcription Factors.

References :

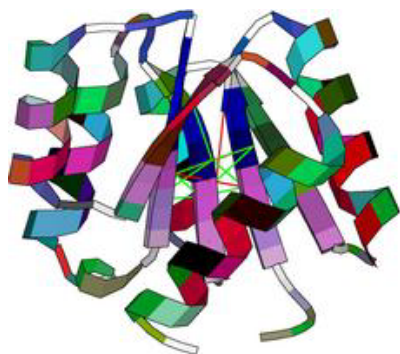
- ¹Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2001) An overview of the structures of protein-DNA complexes. *Gen. Biol.*, 1, 1.
- ²Harrison, S.C. (1991), A structural taxonomy of DNA-binding domains, *Nature*, 353, 715-719.
- ³Luscombe, N.M. and Thornton, J.M. (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, 320, 991-1009.
- ⁴Boggon, T.J., Wei-Song Shan, Santagata, S., Myers, S.C., Shapiro, L., (1999), Implication of Tubby Proteins as Transcription Factors by Structure-Based Functional Analysis, *Science*, 286, 2119-2125.
- ⁵Jones, S., Shanahan, H.P., Berman, H.M., and Thornton, J.M., (2003), Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins, *Nucl. Acid. Res.*, 31, 7189-7198.
- ⁶Stawiski, E.W., Gregoret, L.M., Mandel-Gutfreund, Y., (2003), Annotating Nucleic Acid-Binding Function Based on Protein Structure, *J.Mol.Biol.*, 326, 1065-1079.

Number: 40

Authors: Robert M. MacCallum¹

¹Stockholm Bioinformatics Center, maccallr@sbc.su.se

Title: Self-Organized Maps of Sequence Profiles for Protein Structure Prediction



Short abstract: Self-organizing maps have been shown to be useful for visualising the high-dimensional data associated with sequence profile windows and for contact prediction (MacCallum, R.M., ISMB/ECCB 2004 Long paper #45). This laptop session provides an ideal opportunity to discuss the meaning of these results while looking at 3D visualisations in real time.

Full abstract: Sequence profiles, such as those produced by PSI-BLAST, have been widely used in sequence-structure studies and prediction. This is because they summarise the amino acid preferences at each (equivalent) sequence position for a whole family of proteins. The sequences we see today evolved within the constraints of the overall 3D structure of proteins and also their function (catalytic sites and interactions with other molecules). Profiles therefore contain valuable information about structure and function. Secondary structure prediction methods, such as Ph.D. and PSIPRED, are trained on windows of profiles to predict (typically) three structural states: helix, strand and coil. At CASP5 there was some debate about the use of other and perhaps more informative structural "alphabets" (see also Karchin et al, 2003) as the target for local structure prediction methods and for subsequent use in remote-homology detection. Some time ago, Han and Baker (1996), took the alternative approach of clustering sequence profiles and identifying clusters with consistent structures. This has since been developed further into the I-sites library by Chris Bystroff. I-sites contains only the highly correlated sequence clusters however, and its purpose is not sequence profile visualisation.

In my paper to be presented at ISMB this year (MacCallum, 2004) I show that Kohonen's self-organising map (SOM) is a useful tool for sequence profile window clustering. Using a 3D SOM implementation (most SOM grids are 2D), the mapping can be easily converted into RGB colours for display on a 3D structure. Mappings using profile windows provide a more informative "view" of sequence-structure information than single profile columns, and the most noticeable result is the striping seen frequently across parallel sheets (see Representative Figure). Anti-parallel strands also show occasional anti-parallel striping. Edge strands, not surprisingly, exhibit different colour patterns to non-edge strands. In the paper, I show how these observations can be exploited in a contact prediction method which performs well without the need for correlated mutation information. In this laptop session I invite you to look at structures with this new visualisation method and discuss its more general meaning and purpose. For those who are interested I will briefly explain an approach to order-disorder prediction which also uses profile window SOMs.

Representative Figure description: The CheY protein (PDB code 3CHY) is shown using a colour scheme derived from a SOM mapping of 15-residue sequence profile windows. Note the striping in the parallel sheet and the edge strand discrepancy. Predicted contacts are indicated with lines: green are correct (<8 Å C-beta separation), red are incorrect.

References

Han KF, Baker D. "Global properties of the mapping between local amino acid sequence and local structure in proteins." *Proc Natl Acad Sci U S A*. 1996 Jun 11;93(12):5814-8.

Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. "Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry." *Proteins*. 2003 Jun 1;51 4):504-14.

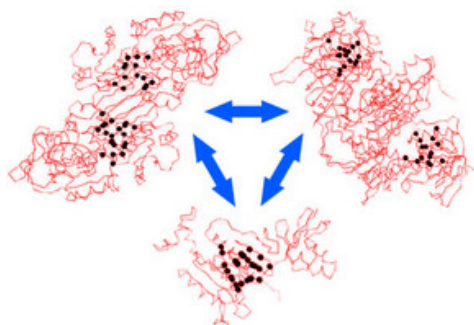
MacCallum, RM. "Striped Sheets and Protein Contact Prediction." ISMB/ECCB 2004 Long paper #45.

Number: 42

Authors: Craig J. Lucas¹, Andrew Bulpitt²

University of Leeds, ¹craigl@comp.leeds.ac.uk ²andyb@comp.leeds.ac.uk

Title: Automatic Identification of Key Patterns Within Multiple Protein Structures



Short abstract: We present an algorithm that searches for multiple key patterns between protein structures, optimised for comparing multiple structures with common function but different fold. We demonstrate the algorithm running on three proteins, showing that finding patterns by comparing all three together is preferable to comparing two of the three separately.

Full abstract: Introduction

We present an algorithm with two main goals. Firstly, an efficient way of searching for multiple, key features between protein structures as opposed to scoring an overall match. Secondly, to optimise the comparison of multiple input structures sharing a common function. As many similar structures as are present may be found, rather than producing one largest match. There is no backbone limitation and no limit on matched pattern size. We demonstrate the algorithm running on three adenine-binding proteins, finding their active sites as their common sub-structures and showing that comparing all three structures together is only marginally slower than comparing any two of the three.

Method

The algorithm is a graph-based, breadth-first search, starting by comparing the smallest patterns of only two residues or atoms, depending on the desired level of matching. A match occurs if every inter-node distance of a pattern is within a given tolerance of a corresponding distance in the matched pattern. The tolerance is empirically defined, based on known error in the input data.

Any matches across the entire set of input structures are repeatedly expanded and searched until no larger common patterns remain.

When multiple proteins of differing overall structure with small sections in common are given as input, the algorithm removes parts of the structure with no common features early on. The more structures that are input, the lower the probability that small structures will match across the entire set by chance. This means it is likely that a final result will be obtained faster than by comparing proteins of a set two at a time. The search complexity may be further reduced by a coherence value which forces each point within a pattern match to lie within the coherence distance from at least one other point within the same pattern. This allows a match to be of any size, but not split over a large distance in the found structure.

Results

To demonstrate the algorithm, the example here is a comparison between three proteins - an NAD-binding alcohol dehydrogenase (PDB code 1HDX), an FAD-binding trypanothione reductase (PDB code 1AOG) and an ATP-binding moeb-moad protein complex (PDB code 1JWA). These proteins bind different ligands, but each ligand contains adenine, providing an example of multiple structures which are likely to contain similarities but which differ in overall backbone structure.

The algorithm was run with a matching tolerance of 3 angstroms and a coherence value of 6 angstroms.

No limit was placed on the diameter of the pattern to be found. Only carbon alpha atoms were considered, to limit the complexity of the search.

No limitations were put on the regions of the input structures to consider, resulting in input structures of 748 nodes, 485 nodes and 298 nodes.

Figure 1

illustrates the matching atoms as circles, superimposed at their locations within the associated backbone structure. The largest pattern found to match between all three proteins consists of 14 nodes.

The algorithm was run on a 2.4Ghz Intel Zeon processor, with 1GB of memory.

Table 1 shows the time taken for the algorithm to finish running on each combination of two structures from the three, and with all three considered at once.



Figure 1: Three adenine-binding protein structures, with highlighted points showing the location of the largest matches found.

1HDX vs 1JWA	297 seconds
1HDX vs 1AOG	> 1 hour
1JWA vs 1AOG	> 1 hour
1HDX vs 1JWA vs 1AOG	325 seconds

Table 1: Time taken for algorithm to complete for each combination of inputs.

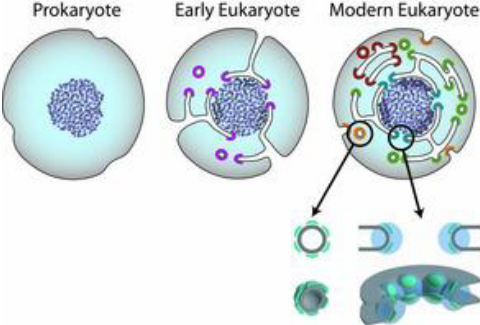
Discussion

The two matching regions found within the NAD-binding protein are at ligand-binding sites, as are the two found within the FAD-binding protein and the matching region found within the asymmetric unit of the ATP-binding protein. As the timings in Table 1 illustrate, the comparison of all three proteins together takes 28 seconds longer to complete than the fastest comparison of two structures, between 1HDX and 1JWA. Comparisons between 1HDX vs 1AOG and 1JWA vs 1AOG took considerably longer due to the process running low on memory. These comparisons failed to reduce the number of common patterns early enough to keep the search space low, demonstrating another advantage to comparing all three simultaneously. Although some results, such as those shown here, have been obtained by running this algorithm on a single processor, when scaled to larger problems this appears to become infeasible. The algorithm is well suited to parallel implementation, however and may be split between multiple processors. Also, the algorithm is only required to run once only for each set of structures with some shared function.

Conclusion

We have presented an algorithm which can find multiple common features within protein structures, unlimited by size and with full structures as input. It is especially encouraging that the algorithm identified both separate active sites within the NAD-binding and FAD-binding proteins, rather than just a single largest common match. Future work will build on these initial results to develop a parallel implementation, capable of processing multiple structures simultaneously with the ultimate intention of seeking common features at the atom level, not just at the residue level.

Number: 44
Authors: Damien Devos¹, Svetlana Dokudovskaya², Marc A. Marti-Renom, Rosemary Williams, Brian T. Chait, Andrej Sali, Michael P. Rout
¹UCSF, damien@salilab.org ²Rockefeller U.
Title: Evidence for A Common Evolutionary Origin of Nuclear Pore Complexes and Coated Vesicles



Short abstract: We report a structural analysis of the seven proteins that comprise the yNup84/vNup107 subcomplex. Our results suggest a common evolutionary origin for nuclear pore complexes and coated vesicles in an early membrane-curving module that led to the formation of the internal membrane systems in modern eukaryotes.

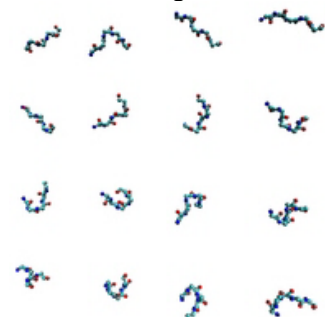
Full abstract: Numerous features distinguish prokaryotes from eukaryotes, chief among which are the distinctive internal membrane systems of eukaryotes. These membrane systems form elaborate compartments and vesicular trafficking pathways, and sequester the chromatin within the nuclear envelope. The nuclear pore complex is the portal that specifically mediates macromolecular trafficking across the nuclear envelope. It is not clear how the nuclear pore complex could have evolved from organisms with no analogous transport system. The yNup84/vNup107 subcomplex is a core building block of the nuclear pore complex. Here we report a structural analysis of the seven proteins that comprise this subcomplex. Our analysis reveals that all seven proteins contain either a beta-propeller or an alpha-solenoid or both. In addition, these folds are shared with major components of the vesicle coating complexes associated with vesicular trafficking. These similarities suggest a common evolutionary origin for nuclear pore complexes and coated vesicles in an early membrane-curving module that led

Number: 48

Authors: Cristina Benros¹, Alexandre G. de Brevern², Serge Hazout

¹EBGM, INSERM E0346, benros@ebgm.jussieu.fr ²debrev@ebgm.jussieu.fr

Title: Predicting Local Structural Candidates from Sequence by the "Hybrid Protein Model" Approach



Short abstract: We are developing a structural profile-based method for local protein structure prediction from sequence. The aim is to propose long local structural candidates. The prediction scheme is based on a novel clustering method, named "Hybrid Protein Model". The principle of our strategy is to carry out a sequence–structure alignment.

Full abstract: Predicting protein structure from amino acid sequence constitutes a major challenge. Both X-ray crystallography and nuclear magnetic resonance analysis present technical limitations. Hence, prediction methods constitute an alternative and will contribute to better understand proteins biological functions.

We are developing an innovating and efficient computational method for local protein structure prediction from amino acid sequence. We assume that protein local structural information is encoded to a large extent in local amino acid sequences. Our aim is to propose long local structural candidates for protein sequences presenting poor sequence identity with known protein structures.

Our local protein structure prediction scheme is based on (i) a simplified description of protein 3D structures by means of a structural alphabet identified in a previous work, i.e. a set of 16 short structural fragments of 5 residues length called Protein Blocks (PBs) able to accurately approximate protein structures (see Figure ; de Brevern et al., 2000, Fourier et al., 2004), and (ii) a novel structural clustering method, named Hybrid Protein Model (HPM; de Brevern and Hazout, 2001, 2003), which extends to longer fragments the local structure description. HPM is a circular self-organizing neural network close to Self-Organizing Maps.

The process is composed of three steps.

(i) Learning of proteins local structural conformations:

The HPM method is used to learn protein structure fragments encoded into the structural alphabet (fragments of 7 PBs, i.e. 11 amino acids). HPM performs a multiple structural alignment of the fragments on the basis of the similarity of the PB series, and by this way enables the compression of the structural information into a library of overlapping clusters grouping structurally similar fragments. Each cluster is represented by an average local structural prototype. The whole of the prototypes is representative of protein local conformations. The library obtained is composed of 120 overlapping prototypes showing a good structural stability (mean *rmsd*, root mean square deviation, of 0.95 angstrom) and a high continuity between the successive local folds.

The main interest of the method is to represent the library as a "structural profile", i.e. a matrix of 120 PB probability distributions (i.e. 16x120). Hence, the matrix elements give the frequency of each PB at each site of the multiple structural alignment. Seven successive probability distributions characterize a structural prototype of the library and two successive prototypes, which are overlapping, have in common 6 probability distributions. Hence, HPM defines a probabilistic protein. The HPM method has been exhaustively evaluated according to the different parameters used. In addition, an improvement of the training has been carried out (Benros et al., 2003).

(ii) Analysis of the sequence-structure relationship:

The investigation of the sequence-structure relationship has shown a high amino acid information content of the structural prototype library. Specific relations and sequence preferences have been highlighted through the computation of amino acid occurrence matrices. Classical preferences appear clearly (e.g. Alanine and Leucine in alpha-helices, Isoleucine and Valine in beta-strands, Glycine and Asparagine in turn regions) and more unusual relations are pointed out.

(iii) Proposition of local structural candidates:

Currently, we are optimizing a structural profile-based method for predicting local protein structure. The principle of our strategy is to carry out a sequence - structure alignment by computing a scoring matrix which measures the local compatibility between a target sequence window and the different regions of the structural profile built by the HPM method. These sequence - structure adequacy scores are computed by using the amino acid occurrence matrices defined previously. A dynamic programming approach enables us to locate and extract the best scoring local regions in the profile. These regions characterize particular overlapping structural prototypes. This strategy allows to propose for the given sequence window, several structural candidates showing the same sequence-structure adequacy. Despite the use of a fixed fragment length for the training, the HPM property of continuity due to the overlap allows the definition of longer structural regions.

When only one candidate is proposed for each overlapping sequence window, the results obtained are already promising. The local prediction rate in terms of PBs, noted Q16, which corresponds to the ratio between the number of PBs correctly predicted and all the PBs is equal to 43.8%. Using the *mda* measure (maximum deviation in backbone torsion angles; Bystroff and Baker, 1998), 47.2% of the residues fall into at least one eleven-residue fragment with no backbone angle deviation greater than 120 degrees. This percentage reaches 57.3% with eight-residue fragments.

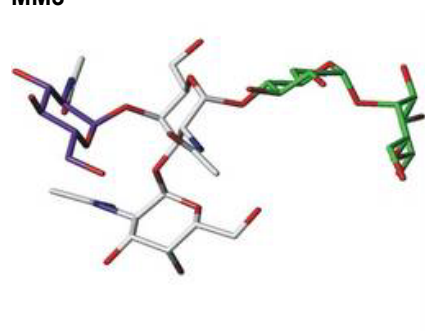
The sequence-structure relationship is currently optimized thanks to a strategy called "Sequence Families" which takes into account the fact that the mapping between sequence and structure is degenerate in order to improve the local structure prediction. Locally, the other potential candidates are also analyzed. A clustering of the different candidates according to their structural similarity in terms of *rmsd* will enable us to take account of the local conformational variability especially in regions where the sequence-structure relationship remains fuzzy. As a perspective, the local structural candidates proposed could be used as structural constraints in *ab initio* global structure modelling. The global structure could also be reconstructed by combinatorial assembly of these candidates.

References :

- Benros C, de Brevern AG, Hazout S. Hybrid Protein Model (HPM): A method for building a library of overlapping local structural prototypes. Sensitivity study and improvements of the training. IEEE Int Work 2003;53-70.
 Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motif. J Mol Biol 1998;281:565-77.
 de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41:271-287.
 de Brevern AG, Hazout S. Compacting local protein folds with a Hybrid Protein. Theor Chem Acc 2001;106(1/2):36-47.
 de Brevern AG, Hazout S. Hybrid Protein Model for optimally defining 3D protein structure fragments. Bioinformatics 2003;19:345-353.
 Fourrier L., Benros C. and de Brevern A.G. (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures, BMC Bioinformatics, 5, 58.

Number: 49

Authors: **Abraham Nahmany¹, Francesco Strino, Jimmy Rosen, Graham J.L. Kemp, Per-Georg Nyholm**
 Department of Medical Biochemistry, Gothenburg University, ¹avi_nahmany@yahoo.se
 Title: **Prediction of Oligosaccharide Conformations using Genetic Algorithms and Molecular Mechanics MM3**



Short abstract: In the present study, we have implemented a system called GLYGAL for the prediction of oligosaccharides 3D structures. GLYGAL performs conformational searches using a parallel genetic algorithm. The searches are performed in the torsion angle conformational space and energy calculations are performed using the MM3(96) force field.

Full abstract: In predicting oligosaccharide conformations, achieving a good trade-off between good sampling of the conformational space and computation time is a major problem. For example, sampling a pentasaccharide with eight torsion angles in 15° steps around each rotatable bond results in over 100 thousand million conformations, and it would take years of CPU time in order to find the conformation with the lowest energy. Genetic algorithms (GAs) have been shown to achieve a good trade-off in similar applications, so the aim of this work was to investigate whether GAs can be used efficiently in modelling oligosaccharides.

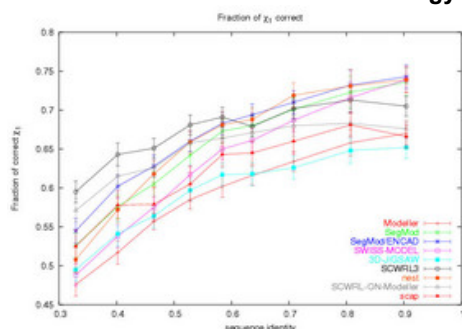
In this study, we have implemented a system called GLYGAL that can perform conformational searches using genetic algorithm (GA) search methods. The searches are performed in the torsion angle conformational space and energy calculations are performed using the molecular mechanics, MM3(96) force field. The system includes a graphical user interface for setting calculation parameters and incorporates a 3D molecular viewer. The results obtained using different GAs implemented in GLYGAL were compared with each other, and with results obtained from MM3 filtered systematic search. Using GLYGAL we were able to investigate different oligosaccharide structures showing that genetic algorithm search methods, though based on heuristics, are a very good tool for the task in hand. We also show that using GAs is faster; e.g. in the case of a pentasaccharide, at least fifteen times fewer conformations need to be sampled in comparison to the filtered systematic search.

Number: 50

Authors: Björn Wallner¹, Arne Elofsson²

Stockholm Bioinformatics Center, ¹bjorn@sbc.su.se ²arne@sbc.su.se

Title: Benchmark of Different Homology Modelling Programs



Short abstract: In this study we have benchmarked freely available homology modelling programs: Modeller, SWISS-MODEL, SegMod/ENCAD, 3D-JIGSAW and *nest*. We also included two methods for predicting side-chains: SCWRL and *scap*. All methods were given the same alignment and the overall quality and stereochemistry of the resulting models were analyzed.

Full abstract: The most successful methods for predicting protein structure, predict the structure of a protein sequence based on its homology to a sequence with known structure. This is done by aligning the sequence of the unknown and the known structure. To get the structure right it is important to have the correct alignment. But once you have an alignment you can feed that alignment to a homology modelling program and the program will use the information in the alignment to build a 3D model of the protein. There are a number of difficulties involved in this step like how should deletions and insertions be modelled? and how do to position the side-chains? In this study, we have used alignments between protein domains belonging to the same SCOP family (with sequence identity ranging from 30%-100%) as an input to five different homology modelling programs, Modeller, SegMod/ENCAD, SWISS-MODEL, 3D-JIGSAW and *nest* within the JACKAL modeling package. As a further reference SCWRL3 and *scap* (also within the Jackal modeling package) were used to build side-chain from simple backbone models.

In general there is not a huge difference between the different methods. But looking at the details there are differences. The differences are most pronounced for the side-chain prediction. For backbone dihedrals all methods perform equal except SegMod/ENCAD which has 10 percentage points lower fraction of phi/psi dihedrals correct. However these models have good stereochemistry which indicates that it is difficult to get both the correct stereochemistry and correct backbone dihedrals.

The most interesting result is that SCWRL3 clearly predicts side-chains better than any other method for sequence identities below 50%, while *nest*, SegMod/ENCAD and SWISS-MODEL all perform better than SCWRL3 for high sequence identity (>80%). This suggest that modelling seem to help in side-chain prediction for close homologs, while it is not so helpful for more distant homologs. SCWRL3 superiority for more distant homologs is probably because it is using a backbone dependent rotamer library, and at low sequence identity it is more favorable to choose the most probable side-chain rotamer keeping a fixed backbone, than try to model both the backbone and side-chain at the same time.

It has been shown in many studies and also at CASP5 that a model very seldom is more close to the native structure than the template it was build on. This is also true for most cases in this benchmark. However *nest* seem to make a larger fraction of the models better than worse, 17% of all models has 0.01 better MaxSub than the template, while only 8% are worse than 0.01. This improvement is tiny, but still it is clear that *nest* rarely makes the models worse than the template. Modeller and SWISS-MODEL makes the largest changes from the template.

Number: 51

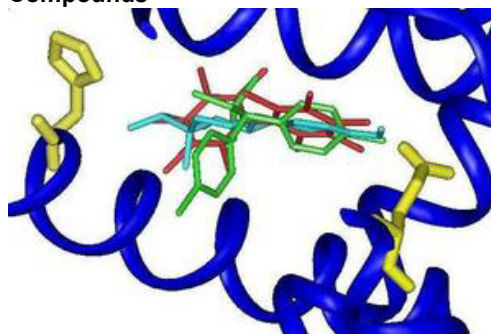
Authors: Patrick May¹, Thomas Steinke², Michael Meyer

Zuse Institute Berlin, ¹patrick.may@zib.de ²steinke@zib.de

[illegible]

Full abstract: THESEUS is a parallel implementation of a protein threading based on a branch-and-bound search algorithm to find the optimal threading through a library of template structures. The template fold library is built on SCOP (Murzin et al. 1995) domains, which are available as ASTRAL PDB-style files (Brenner et al. 2000). THESEUS uses a template core model based on secondary structure definition and a scoring function based on pseudo energies that includes pairwise contacts, solvent accessibility, homology and variable gap lengths. The threading core is implemented in C++ and designed as a SPMD parallelization architecture using MPI for parallel communication. Initial benchmarks show a quite well scaling up to 32 nodes on a 16 node dual CPU Linux cluster. A validation of the threading results has been done on the basis of the Fischer benchmark (Fischer et al. 1996), which uses correct scop fold determination as validation criterion. The results show, that the current threading implementation performs quite well on alpha only, beta only, and alpha-beta domains, but have some difficulties on alpha-and-beta domains. The THESEUS threading core is constructed to be part of an automatic protein structure prediction environment based on Grid and Web Service technology which is under development at ZIB Berlin.

Number: 52
 Authors: **Ermanna Rovida¹, Pasqualina D'Ursi², Paola Fossa, Luciano Milanesi**
 Institute for Biomedical Technologies - CNR, ¹ermanna.rovida@itb.cnr.it ²pasqualina.dursi@itb.cnr.it
 Title: **Modelling the Interaction of Human Estrogen Receptor alpha with Organic Polychlorinated Compounds**



Full abstract: The organic polychlorinated compounds like dichlorodiphenyltrichloroethane (DDT) with its metabolites and polychlorinated biphenyls (PCBs) are present in atmospheric particulate as persistent contaminants. They have been recognized to have detrimental health effects both on wildlife and humans acting as endocrine disrupters (EDC). This effect is obtained by mimicking the action of the hormone estrogen thus interfering with hormone response at different points. There are several experimental evidences that they bind and activate estrogen receptors (Kuiper GG et al., 1988, *Endocrinology* 139: 4252-63). Despite the growing concern about the toxicological activity of EDC, molecular data of the interaction of this compounds with biological targets are still lacking.

64

were calculated. Docking simulations were performed with Autodock 3.05 that allows to take into account the ligand flexibility (Morris, G. M., et al. 1998, J.Comp.Chem., 19: 1639-1662) using the LGA algorithm method with a blind text procedure which explores the surface of the protein with a grid encompassing the whole molecule. We obtained a good reproducibility of the estradiol binding with the crystallographic complex (RMSD 0.6). Further refinement were obtained by use of the program QXP (McMartin C, Bohacek RS. 1997, J.Comp.Aid.Mol.Des., 11:333-344) that includes receptor side chains flexibility.

Docking simulations generated several ligand conformations that were ranked on the basis of the docking energy parameter calculated by the program. In all cases the most favourable docking energies corresponded to those conformations located in the estradiol binding pocket. The affinity values estimated by Autodock (1/ki) were comparable for all ligands and were one order of magnitude lower than estradiol.

To better define the aminoacidic interactions, the docking simulations were repeated with Autodock and with QXP restricting the surface of interaction to the receptor binding pocket. For p,p-DDT, three clusters of conformations were obtained that differed for the orientation of the C-Cl3 group while for the other ligands only one cluster was found. In all cases, residues involved in short range interactions are mainly hydrophobic confirming the blind test results. None of the hydrogen bonds stabilizing the estradiol binding are conserved except for a hydroxylated PCB metabolite (PCB-OH).

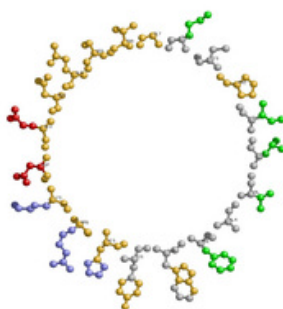
Our results suggest a competitive binding of EDC to Estrogen receptor alpha and a possible synergistic effect. The EDC binding sites characterized in this model will be used to explore the protein database to search for other possible biological targets of these compounds.

Number: 54

Authors: **Julian Mintseris¹, Zhiping Weng²**

Boston University, ¹julianm@bu.edu ²zhiping@bu.edu

Title: **Optimized Protein Representations from Information Theory**



Short abstract: We address the long-standing problem of representing a protein by an alphabet of functionally similar atom/residue groups. Using information theory we can obtain an optimized protein representation from protein monomer or protein interface datasets that are in agreement with each other and with general concepts of protein energetics.

Full abstract: The problem of describing a protein representation by breaking up the amino acids atoms into functionally similar atom groups has been addressed by many researchers in the past 25 years. They have used a variety of physical, chemical and biological criteria of varying degrees of rigor to essentially impose our understanding of protein structures onto various atom-typing schemes used in studies of protein folding, protein-protein and protein-ligand interactions, and others. Here, instead, we have chosen to rely primarily on the data and use information-theoretic techniques to dissect it. We show that we can obtain an optimized protein representation for a given alphabet size from protein monomers or protein interface datasets that are in agreement with general concepts of protein energetics. Closer inspection of the atom partitions led to interesting observations pointing to the greater importance of the hydrophobic interactions in protein monomers compared to interfaces and, conversely, greater importance of polar/charged interaction in protein interfaces. Comparing the atom partitions from the two datasets we show that the two are strikingly similar at alphabet size of five, proving that despite some differences, the general energetic concepts are very similar for folding and binding. Our methods will be useful in defining atom types for small molecule ligands in protein-ligand interaction and QSAR studies as well as in identifying protein structural motifs. In addition, our approach provides a convenient way to visualize and a concise way to describe basic concepts of protein structure and interaction energetics.

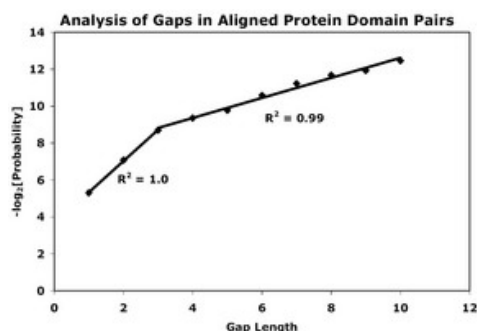
Number: 55

Authors: **Nalin C. Goonesekere¹, Byungkook Lee²**

Bioinformatics & molecular modeling section, Laboratory for Molecular Biology, National Cancer Institute,

¹goonesen@pop.nci.nih.gov ²bk@nih.gov

Title: **Frequency of Gaps Observed in a Structurally Aligned Protein Pair Database Suggests a SIMPLE GAP PENALTY FUNCTION**



Short abstract: The logarithm of the frequency of gaps observed in a database of structurally aligned protein domain pairs is found to vary linearly with the length of the gap, but with a break at a gap of length 3. These results suggest a modification of the affine gap penalty function.

Full abstract: Homology search tools generally use a score function that assigns a score for each aligned residue pair and a penalty for each gap in the alignment. The score for aligned residue pairs is obtained from examining large numbers of aligned protein sequences and there is a firm theoretical basis that provides guidance on how to select the score function that will yield an optimal alignment. The gap penalty is also important for the sensitivity (ability to find remotely related sequences) of the search tool and for the accuracy of the alignment. Most sequence homology detection algorithms employ the 'affine gap penalty' scheme of the form $(q + r \cdot n)$, where n is the length of the gap, and q and r are empirically chosen parameters representing the cost of opening and extending a gap, respectively. The affine gap penalty scheme recognizes that it costs to open a gap as well as to extend an existing one, and has proved to be superior to length proportional gap costs, which often produce a large number of short insertions or deletions. However, no firm theoretical or experimental support has been presented for such a gap penalty scheme and it has also been criticized for over-penalizing long gaps.

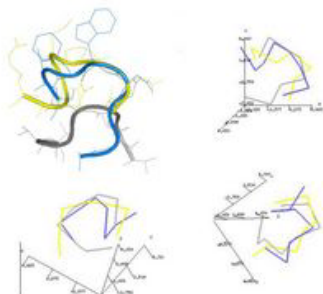
We examined the pattern of gaps that occur in a database of structurally aligned protein domain pairs. We find that the logarithm of the frequency of gaps varies linearly with the length of the gap, but with a break at a gap of length 3, and is well approximated by two linear regression lines with R^2 values of 1.0 and 0.99. The bilinear behavior is retained when gaps are categorized by secondary structures of the two residues flanking the gap. Since the bilinear behavior is observed for all secondary structural environments, the mechanism that produces this behavior is operating outside of the constraints of the secondary structure of the protein. Similar results were obtained when another, totally independent, structurally aligned protein pair database was used. These results suggest a modification of the affine gap penalty function, to a bilinear gap penalty function. The bilinear gap penalty function is a special case of a piece-wise linear gap penalty made of K straight lines, where $K = 2$. The computational complexity for a bilinear gap penalty function has been shown to be $O(mn)$, the same as for the affine gap penalty, indicating that computational cost would not be a factor in the implementation of such a function in homology detection algorithms. Though the logarithm of the probability distribution of gap lengths for all gap types exhibit the same bilinear behavior, the parameters describing the regression lines are clearly a function of the secondary structure of residues flanking the gap. Such secondary structure-dependent gap penalty functions could be useful in cases wherein protein sequences are aligned to a known protein structure.

Number: 56

Authors: Victoria F. Dominguez Del Angel¹, Luis Mochan², Bernard Caudron, Charles Roth

¹bbmi - Pole Informatique; Institut Pasteur, victoria@pasteur.fr ²Centro de Ciencias Fisicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Mexico, mochan@em.fis.unam.mx

Title: PP2A(alpha) Interactions Domains, a Model of 3D Pattern Representation



Short abstract: The Protein phosphatase-2A is a holoenzyme with pleiotropic functions. Based on analysis in amino-acid composition and 3D representation, we developed a in-silico approach that detects small fragments in proteins that might interact with AC core of this protein phosphatase. This strategie could be extended to detect small binding fragments in other proteins.

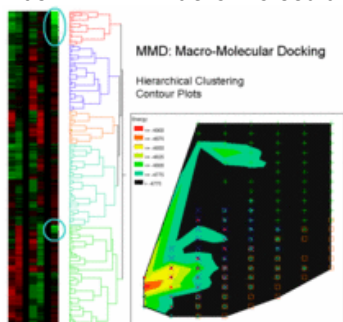
Full abstract: Protein phosphatase 2A (PP2A) is one of the major classes of serine/threonine phosphatases that is critical to many cellular processes. The core structure of PP2A consists of a catalytic subunit (PP2Ac) and scaffold protein (A/PR65). This core dimer (AC) can recruit a third regulatory B subunit that determines the substrate specific for the enzyme. The core has also been shown to interact with other proteins, including T antigens from virus polyomavirus and SV40, casein kinase 2 and chemokine receptor CXCR2. Little is known about how these different proteins bind to the core. Based on analysis of the sequences and structure, we propose a method for representing and detecting three dimensional patterns of side chains in protein structures. We normalize the representation by shifting the center of mass of the domain to the origin of the coordinate system, and by rotating the domain so that its principal inertia axes coincide with the cartesian axes. The normalized representation allows a comparison between the 3D shape of different domains.

Number: 57

Authors: Qingjuan Gu¹, William W. Reenstra, John J. Rux²

The Wistar Institute, ¹qingjuan@wistar.upenn.edu ²rux@wistar.upenn.edu

Title: MMD: A Macro-Molecular Docking Program for Locating Domain-Domain Interactions



Short abstract: MMD is a macro-molecular docking program designed to take advantage of distributed computing resources so that multiple docking solutions can be evaluated in parallel on a wide range of hardware. The ready availability of inexpensive PC's and open-source software makes these types of studies feasible on a limited budget.

Full abstract: The activation and regulation of biological macromolecules commonly involve specific interactions between individual domains where ligand-dependent conformational changes in one domain are coupled to conformational changes in a second domain. MMD is a new macromolecular docking program specifically developed to establish sites of domain-domain interaction within flexible molecules. The MMD program evaluates a series of docking orientations for each macromolecular complex in parallel. Molecular motions are simulated with the molecular dynamics program NAMD and a CHARMM force field to allow for conformational flexibility and to minimize the energy of each complex. Symmetry operators can be applied to accommodate known biological properties (e.g. 2-fold symmetric dimer) and to limit the size of the search space (fewer degrees of freedom).

For maximum portability the application is written in the Java language and uses CORBA (Common Object Request Broker Architecture) to implement multiple processor computation in parallel. CORBA is a standard architecture for distributed object systems (object-oriented). It eases the transfer of information between different machines by allowing a distributed, heterogeneous collection of objects to interoperate on multiple operating systems. MMD enables a heterogeneous mixture of unix and linux machines to operate as a cluster with a master/slave architecture. The master node uses the multi-thread feature of Java to launch jobs containing a series of docking orientations on each of the slave nodes. The master node also monitors the status of the slave nodes and implements load balancing. The slave nodes simultaneously calculate structures for each docked complex, initiate the molecular dynamics simulation, and finally return the docked structure and energy results back to the master node. The master node tracks and organizes the results until the list of orientations has been exhausted. The resulting docked conformations are sorted by energy and the lowest energy conformations are clustered into groups of spatially-related solutions.

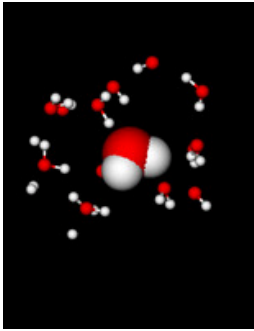
We have specifically designed the MMD program to take advantage of distributed computing resources so that multiple docking solutions can be evaluated in parallel on a wide range of hardware. Typically, a small cluster of Linux PCs are required to run continuously to support the molecular mechanics and docking calculations. The ready availability of inexpensive commodity PC's and open-source software makes these types of studies feasible on a limited budget.

Number: 58

Authors: Gianni De Fabritiis¹, Rafael Delgado-Buscalioni, P. V. Coveney

University College of London, ¹g.defabritiis@ucl.ac.uk

Title: A Minimisation Method for the efficient Insertion of Solvent Water Molecules



Short abstract: Many molecular dynamic simulations involving open systems require the insertion of solvent water molecules. This work provides a fast algorithm for on-the-fly insertion of water molecules.

Full abstract: The insertion algorithm places the incoming molecules at sites where the energy equals a prescribed "target energy". The insertion energy equals the chemical potential of the system in a way that the incoming molecules have a small effect on the local structure of the liquid. The insertions are local in the sense that they are made within the nearest available volume with respect to the initially proposed insertion site. The present algorithm builds on the ideas behind the usher insertion algorithm for Lennard-Jones atoms and generalise it to polyatomic and polar solvents. While the insertion is being performed the system's configuration is frozen and the incoming molecule is ushered through the potential energy landscape by a sort of multiple steepest descent iterator. The only parameters in the algorithm are the maximum translational and rotational displacements of the usher iterator. These can be extracted from previous inspection of the liquid structure. We consider the TIP3 model for water widely used in biomolecular simulations for which around ten iterations are required to insert a water molecule in the liquid state.

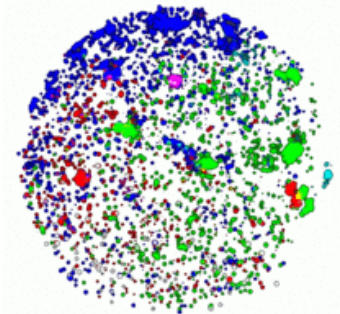
Any minimisation method is efficient only if it reconstructs the energy landscape with a limited number of probes. The main idea of the usher is to optimise the use of the information by searching only inside the energy wells. In fact, any excursion outside the searched well is effectively equivalent to a random restart with the loss of all the information accumulated on the current well. The optimal balance between the amount of details of the well needed to find the target energy and the explorations of new wells is the key passage on which usher is designed.

Number: 59

Authors: Jim Procter¹, Andrew Torda²

¹Centre for Bioinformatics, University of Hamburg, procter@zbh.uni-hamburg.de ²torda@zbh.uni-hamburg.de

Title: Protein sequence-structure Alignments for Prediction, Classification, and Visualization



Short abstract: This poster examines the use of sequence-structure alignments for protein classification, and evaluates the effect that fold conformation and architecture have on the performance of a threading based fold recognition method. The dataset used is also presented as a VRML based visualization available for browsing at <http://www.zbh.uni-hamburg.de/wurst/protospace/>

Full abstract: In protein sequence to structure alignments, or threadings, the aim is to find the optimal arrangement of a protein sequence on some arbitrary protein structure. If the structure has sufficient similarity to the sequence's fold the alignment should both score well and produce a reasonable model for at least part of the sequence. However, as sequence homology decreases, the ideal threading of a sequence onto its structural homologs becomes discontinuous, because of inherent chemical dissimilarity. At this point, perhaps nearing the limit of recognisable fold homology, the assumptions involved in making an efficient sequence-to-structure alignment break down.

Discontinuities in the ideal threading arise where both structures are dissimilar, or some part is non-existent in either template or modelled structure. In practice, these necessary gaps and deletions are rarely detected because spatial interactions between all residues cannot be accurately optimised, and discontinuities between sequence and template are generally penalized. Consequently, alignments between extremely remote homologs

frequently include incorrect mappings, and subsequent analysis of model fitness will often discard an otherwise good solution.

The threading algorithm^[1] used in this work is subject to these unfortunate limitations. All parameters have been numerically optimised, to maximise the structural similarity of threading model and observed structure, and then to rank the best model based on sequence-to-structure fitness. Yet, the use of standard dynamic programming algorithms, and essentially local sequence-to-structure fitness functions at the alignment stage means that it is often impossible to numerically distinguish models with the best possible predictions for some parts of a sequence from those which are well formed, but completely wrong.

In order to improve this situation, a complete set of sequence-structure alignments were made for a significant subset of the PDB using ubiquitous database and distributed processing techniques. These alignments and their models were first analysed to detect consistency between sequence and structure for pairs of molecules providing a means of 'sequence-structure comparison' that is surprisingly robust to differences in contiguous domain architecture. The roughly bimodal distributions for a family of geometric similarity scores were then partitioned to define an empirical confidence measure for structural homology. This measure is combined with raw sequence-to-structure fitness scores to define a protein-protein similarity metric tailored to the specific properties of the threading method, WURST. Amongst other uses, the symmetric distance matrix defined by this sequence-structure metric has been clustered to generate fold groups, and a minimal covering for efficient fold recognition with WURST-like threadings.

All-against-all alignment datasets are necessary for the evaluation of predictive tools, such as the WURST method. The one described above has been used to investigate how WURST's performance is dependent upon the structural class as a protein, which is the subject of a related poster at ECCB/ISMB. The methodology that was applied there is ubiquitous in the literature: receiver-operator-characteristic plots are used to assess the detection of structurally proteins for a sequence amongst a ranked set of WURST sequence-structure alignments. The results of this kind of analysis are often difficult to quantify without examination of specific parts of the dataset, leading to a complex set of cases that properly require detailed discussion. This complexity is inherent, however, because such data sets are high-dimensional structures composed by relations between thousands of different proteins.

The focus of this poster is the effective representation of such complex datasets. Traditionally, planar representations are employed in the classification of proteins, belying the roots of the field in phylogeny and cladistics. Instead, we present a diagram in three-dimensions, allowing a richer depiction of the diversity of protein structure space.

The basis of the depiction is formed by a 3D-embedding of the network of sequence structure relations defined by WURST alignments. This is accomplished with multi-dimensional scaling, which, in combination with the non-linear nature of the similarity metric, yields an aesthetically pleasing spherical distribution. The standard tools of high-performance visualization can now be applied to portray correlations between WURST's view of the protein universe, and other categorisations of protein structure, such as CATH class.^[2] A VRML based visualization of this distribution is available [for browsing](#), and other uses of the distribution, beyond exploring the diversity of protein folds without the use of hierarchical trees, are discussed.

References

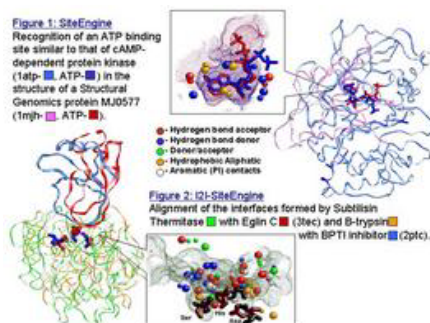
1. WURST: A protein threading server with a structural scoring function, sequence profiles and optimised substitution matrices
Andrew E. Torda, James B. Procter and Thomas Huber *Nucleic Acids Research*, 2004, in press
2. The CATH database: an extended protein family resource for structural and functional genomics. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA *Nucleic Acids Research* 2003, **31**, 452-455

Number: 60

Authors: Alexandra Shulman-Peleg¹, Shira Mintz², Ruth Nussinov, Haim J. Wolfson

¹Tel-Aviv University, School of Computer Science, shulmana@post.tau.ac.il ²Tel-Aviv University, Sackler Inst. of Molecular Medicine, Sackler Faculty of Medicine

Title: Protein Functional Sites Recognition and Classification



Short abstract: We present two novel methods for recognition of protein functional sites. While the first method searches the protein surface for regions similar to known functional sites, the second recognizes similar chemical binding organizations shared by protein-protein interfaces. The methods consider the geometrical and physico-chemical properties and are applicable to large-scale searches of the entire PDB.

Full abstract: Recognition of regions through which protein molecules function and interact is crucial for prediction of molecular interactions which govern most of the cellular processes. Moreover, recognition of such regions can assist in functional annotation of novel proteins. Here we present two methods for recognition of functional sites in protein structures. Unlike methods that compare the locations of the backbone atoms or the identity of the amino acids, the presented methods take into account the physico-chemical properties of both the backbone and the side-chains. Therefore they can recognize similar binding patterns shared by proteins that have no sequence or fold similarity. The first method, SiteEngine, recognizes regions on the surface of one protein that resemble a specific binding site of another. This may suggest the similarity of their binding partners and biological functions. The second method, Interface-to-Interface (I2I)-SiteEngine method, extends this approach to alignment of protein-protein interfaces, which are regions of interaction between two protein molecules. It utilizes the information regarding patterns of interacting chemical groups and performs simultaneous alignment of two pairs of interacting binding sites. I2I-SiteEngine can assist in recognition of certain interface binding organizations that may be important to the formation and stability of the protein-protein complexes. Recognition of such patterns may assist in drug discovery and in prediction of side effects. Both methods are highly efficient and suitable for large scale database searches of the entire PDB.

The input for SiteEngine consists of a binding site of one protein and a complete structure of another protein. The molecular surface shell of the complete molecule is searched for the presence of a region which is similar to the input binding site. The method is based on efficient hashing and matching of triangles of centers of physico-chemical properties and can search large protein structures in a matter of seconds. It investigates every candidate transformation that superimposes at least three centers of physico-chemical properties. We introduce a low-resolution representation by chemically important surface points and efficiently score these solutions to filters out the biologically irrelevant ones. Then, as the number of potential solutions is reduced to a smaller subset, the resolution of the molecular representation is increased, leading to a more precise comparison of the geometrical and physico-chemical properties.

The biological significance of the SiteEngine method is validated on a set of biological applications. First, we introduce a benchmark dataset which is used to construct two databases: one of complete protein structures and the other of binding sites. These databases are used to perform three types of search applications: (1) A given functional site is searched against a large set of complete protein structures; (2) A potential functional site of a protein of interest is compared with known binding sites; (3) A complete protein structure is searched for the presence of an a priori unknown functional site, similar to known sites. While the second application compares between already known binding sites, the first and the third can recognize novel regions that can function as binding sites. Our method is efficient enough to allow computationally demanding applications such as the first and the third. From the biological standpoint, the first and the second applications may identify secondary binding sites of drugs that may lead to side effects. The third application finds new regions that may provide targets for drug design. Each of the three applications may aid in assigning a function and in classification of binding patterns.

In each application SiteEngine has successfully recognized specific types of protein binding sites such as estradiol binding, adenine and ATP binding sites that were used as queries. The same binding sites were further used to search the ASTRAL dataset constructed from the entire PDB. Since SiteEngine searches a complete structure of each protein in a matter of seconds, we find the first application to be the most reliable for such large scale applications. The method was also applied to classification and functional annotation of novel proteins determined as part of the Structural Genomics project. The I2I-SiteEngine method extends the algorithmic approach of SiteEngine to the comparison and alignment between protein-protein interfaces. An interface is defined as an unordered pair of interacting binding sites that belong to two different, non covalently linked, protein molecules. Given two interfaces the goal is to find an alignment that maximizes the similarity between them. This implies solving two problems: defining the correspondence between the binding sites of the two molecules and finding the rigid transformation that will simultaneously maximize the similarity of their corresponding shapes and physico-chemical properties. When considering protein-protein interfaces we are supplied with valuable information regarding the functional groups of two binding sites that interact with each other. Assuming that at least three pairs of interacting functional groups must be present in each interface, we define each such triplet as

an I-triangle. I2I-SiteEngine utilizes this information to perform an even more efficient hashing procedure which is based on I-triangles and allows the simultaneous alignment of the interacting binding sites.

We have used I2I-SiteEngine for the classification of protein-protein interfaces according to their spatial and chemical organizations. Such classifications are important for two types of biological applications. In the first application a 'newly' determined interface is aligned to selected representatives of known binding organizations. In the second application a complete structure of a protein of interest is compared to a database of known interfaces to predict its potential binding partners and binding modes. After using I2I-SiteEngine to classify and search a pilot dataset of some of the common protein-protein interfaces we are currently applying it to classify all the known interfaces stored in the PDB.

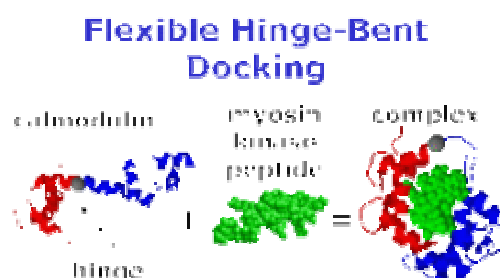
Both of the presented methods can recognize similar physico-chemical patterns shared by evolutionary unrelated proteins. While SiteEngine compares unbound regions on the protein surface, I2I-SiteEngine compares patterns of interaction present in protein-protein complexes. The two methods can be applied synergistically for the full exploration of the available structural data, providing an insight into the mechanism of molecular recognition and protein function.

Number: 61

Authors: Dina Schneidman¹, Yuval Inbar, Ruth Nussinov, Haim J. Wolfson

Tel Aviv University, ¹duhovka@tau.ac.il

Title: FlexDock: An Algorithm for Flexible Hinge-Bent Docking.



Short abstract: We present FlexDock, an algorithm for flexible hinge-bent docking. The algorithm allows hinges placement in one of the docked molecules: receptor or ligand. The number of hinges is not limited. The method performs simultaneous docking of all rigid parts, explicitly considering potential hinge-bending motions.

Full abstract: Introduction

Proteins are highly flexible molecules. It is common to classify protein motions into shear and hinge motion. Shear motions are very limited and involve large number of residues. On the other hand, hinge motions are similar to rotations around an articulated joint and therefore can be very large. Hinge motions involve a small number of residues, since even one bond can provide the required rotational freedom. Other types of motion include allosteric rearrangements, partial refolding, loop and side chain movements. Despite the variety of movements, most of the docking algorithms treat proteins as rigid bodies. Some algorithms are capable of handling some side-chain motions or even loop motions, but modelling large-scale flexibility is a great challenge. There are several approaches to docking of flexible molecules: (i) The rigid subpart method (DesJarlais et al. '86) splits the flexible molecule into subparts, docks each part separately and joins the results into consistent solutions; (ii) The anchor fragment method Rarey et al. '95) selects an "anchor" fragment, docks it and positions sequentially the remaining fragments; (iii) The hinge scoring method (Sandak et al. '95) is based on Geometric Hashing paradigm. It incorporates the information about hinge positions in the initial docking steps. The first two methods can handle only small molecule flexibility.

Here we present FlexDock, an algorithm for flexible hinge-bent docking. FlexDock can place hinges in one of the molecules: receptor or ligand. The number of hinges is not limited. The hinges can be placed at a bond or at a point. The algorithm docks the two molecules simultaneously, considering all possible hinge rotational degrees of freedom and without performing an exhaustive search.

FlexDock Method

The input of FlexDock is two molecules: one rigid molecule and one flexible molecule with a list of hinges. The output is a list of multi-transformations. Each multi-transformation represents one docking solution. Multi-transformation is a list of transformations, one for each rigid part of the flexible molecule. The transformations transform the rigid parts onto the rigid molecule, such that the obtained solution is consistent. The solution of flexible docking is consistent if the hinge points of each rigid part superimpose and there is no intra-penetration between the rigid parts of the flexible molecule.

The FlexDock algorithm has two major stages. The first stage performs rigid docking of all rigid parts. The second stage assembles the rigid docking results into consistent flexible solutions, producing a list of high-scoring multi-transformations. In the first stage we apply the PatchDock algorithm (<http://bioinfo3d.cs.tau.ac.il/PatchDock>) for rigid docking. This algorithm is utilized since it is very efficient and finds the near-native solution in most of the

cases, unless there are serious protein deformations upon binding. The input to the second assembly stage is a list of transformations for each rigid part. We construct an assembly graph, where each transformation is represented by a node. The weight of the node is the score of the transformation. There is an edge between two nodes if their corresponding rigid parts transformations superimpose the hinge point to almost the same position. Source and target nodes are added to the first and last rigid part transformations. The edges are directed from source to target nodes. The weight of the edge is the shape complementarity score between the neighboring rigid parts of the flexible molecule. The list of high-scoring consistent multi-transformations is found by applying the dynamic programming algorithm to the constructed directed acyclic graph.

Experimental Results

The algorithm was tested and verified on a number of cases. One of the cases is the interaction of myosin kinase peptide to calmodulin. Upon binding there is large-scale movement of calmodulin involving splitting of one long helix (see attached figure). The total rotation of one domain relative to the other is upwards of 150 degrees. FlexDock successfully finds the near native result with RMSD 2.12Å ranked second. The corresponding two rigid subpart docking results are ranked only 38 and 6. The running time for this case is 11 minutes. Another test case is the complex of ATP and biotin carboxylase. One of the protein domains is twisted relative to the other domain by 80 degrees upon ATP binding. The rigid subpart docking stage of the algorithm finds the near native results with ranks 161 and 4880. However, the assembly stage succeeds to find the near native complex with rank 17. This is due to the fact that upon binding of ATP, intra-molecular interface is created between the two domains of the protein. The contribution of shape complementarity score of this interface to the flexible result helps to rank it higher.

Discussion

The presented algorithm performs fully flexible hinge-bent docking. There is no restriction to the rotation at the hinge. The algorithm can handle multiple hinges in one of the molecules. The near native result may not rank high for each of the rigid parts, however the combination of the rigid parts scores usually succeeds to increase the rank of the whole complex. The algorithm is very efficient due to several factors: (i) docking of each rigid part is performed only once during the algorithm by a highly efficient rigid docking algorithm based on Geometric Hashing; (ii) the assembly stage performs simultaneous construction of consistent multi-transformations from rigid docking results. The running times of the algorithm are comparable to the running times of rigid docking without hinges. On the one hand we dock smaller parts, reducing the running times, on the other hand the assembly stage is very effective, consuming only few minutes or even seconds. In the future our goal is to add to the algorithm an automatic hinge detection module, that may replace user specified hinge information. In this stage, we assume that there is a contact between each rigid part of the flexible molecule and the rigid molecule. We would like to make the algorithm more robust to handle cases without such a restriction. The greatest challenge is to place hinges in both molecules to perform fully flexible hinge-bent docking.

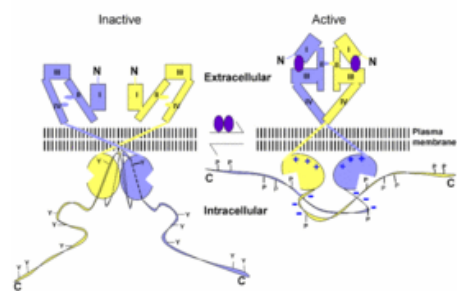
Number: 62

Authors: **Meytal Landau¹**, **Sarel J. Fleishman**, **Nir Ben-Tal²**

Tel-Aviv University, ¹meitalc@post.tau.ac.il ²NirB@tauex.tau.ac.il

Title: **Down-Regulation of the Catalytic Activity of the EGF Receptor via Direct Contact between the Kinase and C-Terminal Domains**

Model of regulation of the EGF receptor



Short abstract: The ErbBs are unique among receptor tyrosine kinases, as the catalytic elements of their kinase domain are constitutively ready for phospho-transfer. The absence of conformational regulation raises a fundamental dilemma: namely, by what mechanism is spurious activation avoided? Our studies, using various computational tools, suggest a novel molecular regulation mechanism.

Full abstract: Tyrosine kinase receptors of the epidermal growth factor receptor (EGFR) family, also known as ErbB or HER, are activated by EGF and other extracellular ligands. These receptors play an important role in the control of many fundamental cellular processes, including cell metabolism, proliferation and differentiation (1). Activating mutations and overexpression of the EGFR family were implicated in a variety of cancers, including carcinoma and glioblastoma. Structurally, the RTKs consist of an N-terminal, extracellular domain that is connected by a short transmembrane span to a tyrosine kinase domain, which is followed by a C-terminal regulatory domain (Figure 1).

The catalytic activity of all receptor tyrosine kinases (RTKs), including the EGFR family, is stimulated in response to appropriate signals, and is encapsulated in multiple layers of control. In most RTKs excluding the EGFR family, ligand-binding to the extracellular domain leads to profound structural changes in the intracellular kinase domain

along with transautophosphorylation of the kinase (2). In contrast, the EGFR family is unique among RTKs in that its activation is independent of its phosphorylation state (3). The X-ray crystal structure of the EGFR kinase domain (4) demonstrated that its unphosphorylated conformation is, in essence, identical to the phosphorylated conformations of other RTKs. The fact that the kinase domain constantly occupies the active conformation leads to an apparent paradox, since it is well known that the EGFRs are not constitutively active (1, 5). Hence, our working hypothesis was that these receptors are regulated by another mechanism intrinsic to the intracellular domain, which is phosphorylation-independent.

The kinase domain forms homodimers in the crystals, where inter-domain contacts are mediated by two segments of EGFR's C-terminal regulatory domain (4). We present here a computational analysis that shows that the complex between the kinase and C-terminal domains is stable and biologically significant. This conclusion is based on a study of the geometric and electrostatic complementarity between the kinases and C-terminal domains within the crystal, and is strengthened by an evolutionary analysis of correlation between amino acid replacements in RTK sequences. Based on our analysis and on experimental data using viral and truncated forms of the receptor, we suggest that the complex between the kinase and C-terminal domains corresponds to the receptor's stable and inactive form, where a substrate is absent. In the inactive conformation (Figure 1; left), each of the extracellular domains assumes a compact structure, and the intracellular domains contact via the C-terminal fragments, leading to an inactive and stable form. Activation (Figure 1; right) occurs when ligands (purple ovals) bind to the extracellular domains, leading to the formation of a stable extracellular contact, which is followed by the rotation of the transmembrane helices (6), and the subsequent destabilization of the contacts between the C-terminal and kinase domains. The kinase can now trans-autophosphorylate the tyrosine residues of its own C-terminal domain as well as tyrosine residues of its protein substrates. Thus, the C-terminal domain serves as a down-regulator of the kinase activity in the EGFR family. Strong reinforcement of this model of regulation is provided by data on EGFR analogues, including human cancer-causing mutants and viral EGFR variants. These EGFR analogues sustained mutations or deletions in the C-terminal fragment contacting the kinase domain, which leads to higher autokinase activity and transforming ability (7). Our model suggests that these mutations destabilize the inactive complex. Our model for the regulation of the EGFR could offer new approaches for the treatment of EGFR-related cancer.

Figure 1

Schematic diagram representing our suggested model of EGFR activation. Two EGFR monomers are colored light purple and yellow. The extracellular domain (residues 1-620, labeled I, II, III, and IV according to its sub-domains) and kinase domain (residues 685-957) are connected via a transmembrane helix (residues 621-642) and a short juxtamembrane segment (not shown). The C-terminal domain comprising 229 amino acids, structure of which has not been determined, follows the kinase domain. Tyrosine (Y) residues, known as the autophosphorylation sites in the C-terminal domain, are indicated. Positive and negative charges are marked in the active conformation on the kinase and C-terminal domains, respectively. In the inactive conformation they roughly neutralize each other.

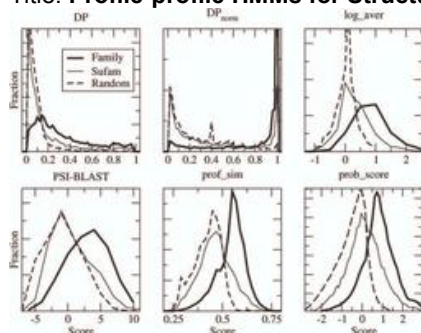
- Schlessinger, J. (2000) Cell 103, 211-25.
- Huse, M. & Kuriyan, J. (2002) Cell 109, 275-82.
- Gotoh, N., Tojo, A., Hino, M., Yazaki, Y. & Shibuya, M. (1992) Biochem. Biophys. Res. Commun. 186, 768-74.
- Stamos, J., Sliwkowski, M. X. & Eigenbrot, C. (2002) J. Biol. Chem. 277, 46265-72.
- Prenzel, N., Fischer, O. M., Streit, S., Hart, S. & Ullrich, A. (2001) Endocr. Relat. Cancer 8, 11-31.
- Fleishman, S. J., Schlessinger, J. & Ben-Tal, N. (2002) Proc. Natl. Acad. Sci. USA 99, 15937-40.
- Boerner, J. L., Danielsen, A. & Maihle, N. J. (2003) Exp. Cell Res. 284, 111-121.

Number: 63

Authors: Håkan Viklund¹, Arne Elofsson²

Stockholm Bioinformatics Center, ¹hakanv@sbc.su.se ²arne@sbc.su.se

Title: **Profile-profile HMMs for Structure Predictions**



Short abstract: In a Hidden Markov model (HMM) the probability that a single sequence is generated by the model is calculated. Here, we show that extending HMMs to calculate the probability that a profile is generated by the model increase the performance both for fold recognition and for membrane topology predictions.

Full abstract: Hidden Markov models have been powerful tools for several bioinformatical problems including, fold recognition, gene finding and topology predictions of membrane proteins. In traditional HMM the probability that a single sequence is generated by the model is calculated. Here, we show that extending HMMs to calculate

the probability that a profile, i.e. a multiple sequence alignment, is generated by the model increase the performance both for fold recognition and for membrane topology predictions. There are several methods the profile information of a query protein can be compared with the multiple sequence alignment information of the HMM, we have tested several of these.

MEMBRANE TOPOLOGY PREDICTION

Methods that predict the topology of helical membrane proteins are standard tools when analyzing any proteome. Therefore, it is important to improve the performance of such methods. Here, we introduce a novel method, PRODIV-TMHMM, which is a profile-based hidden Markov model (HMM) that also incorporates the best features of earlier HMM-methods. In our tests, PRODIV-TMHMM outperforms earlier methods both when evaluated on "low-resolution" topology data and on high-resolution 3D-structures. The results presented here indicate that the topology could be correctly predicted for approximately two thirds of all membrane proteins using PRODIV-TMHMM. The importance of evolutionary information for topology prediction is emphasized by the fact that compared to using single sequences, the performance of PRODIV-TMHMM (as well as two other methods) is increased by approximately 10 percentage units by the use of homologous sequences. On a more general level, we also show that HMM-based (or similar) methods perform superior to methods that focus mainly on identification of the membrane regions.

FOLD-RECOGNITION

The detection of distantly related homologous proteins is one of the basic challenges in bioinformatics. To improve the detection of related proteins it is often useful to include evolutionary information for both the query and target proteins. One method to include this information is by the use of profile-profile alignments, where a profile from the query protein is compared with the profiles from the target proteins. Profile-profile alignments can be implemented in several fundamentally different ways. The similarity between two positions can be calculated using a dot-product, a probabilistic model or an information theoretical measure, see figure. Although the performance of all methods is quite similar, profile-profile methods that use a probabilistic scoring function have an advantage as they can create good alignments and show a good fold recognition capacity using the same gap-penalties, while the other methods need to use different parameters to obtain comparable performances. Another useful method is to use hidden Markov model instead of standard dynamic programming alignment techniques. In this study we combine these two approaches and introduce a novel profile-profile HMM method. We show that this method perform better than fold recognition methods based on either just profile-profile information or HMMs.

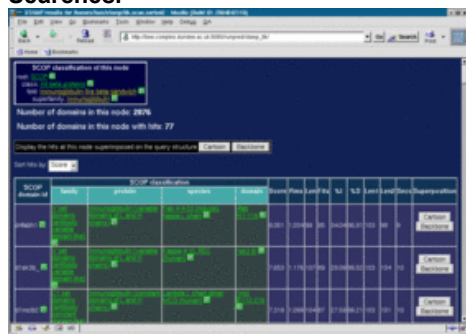
Finally, it should be mentioned that our modular HMM program (modhmm) is freely available under the GPL license.

Number: 64

Authors: Thomas P. Walsh¹, Geoffrey J. Barton²

¹School of Life Sciences, University of Dundee, tom@compbio.dundee.ac.uk ²geoff@compbio.dundee.ac.uk

Title: A PERL Library for Generating Hierarchical, Tree-based Browser Interfaces for SCOP Database Searches.



Short abstract: A Perl library has been developed that creates interfaces for displaying the results of SCOP database searches in a tree-based, hierarchical interface that reflects the SCOP classification. The library is designed to be easily extended to generate interfaces for algorithms of interest.

Full abstract: Advances in computing power make it feasible to search the entire SCOP protein domain database of 54745 domains, using the Map fold recognition algorithm (1) or the STAMP structure comparison program(2), in a reasonable time (< 20 minutes on a 100 CPU Linux cluster). The SCOP classification hierarchy is a significant aid to the interpretation of such database searches. However, visualising the results of such searches can be problematic. Accordingly, a set of Perl classes has been developed that generates a hierarchical, tree-based browser interface for SCOP search results that organises the results to reflect the SCOP hierarchy. The interface allows the user to navigate through the SCOP hierarchy, viewing the hits to the domains in each node of the hierarchy. The library is built on a class that generates a generic interface; interfaces specialised for particular search algorithms are created by subclassing the base class. The basic interface class generates an interface consisting of two windows, one of which displays the SCOP tree as a collapsible menu, while the other displays the hits to domains at a given SCOP node. The node window displays the SCOP

classification of the node and a list of the domains to which there are hits. This window incorporates a sort button that allows the hits at each node to be resorted by various criteria (e.g. score, sequence identity). Subclasses can define lists of sort criteria that are to be associated with the sort button, so that the results for a given algorithm can be sorted by criteria that are appropriate for the algorithm. Resorting of the hits is done by a server-side CGI script that sorts the hits using a sorting function provided by loading classes on the fly; this allows new sorting classes to be easily plugged into the script. The browser interface for fold recognition and structure comparison algorithms can be enhanced by displaying structures in PyMOL (3). Command scripts for PyMOL are generated by a server-side CGI script. As with the sorting script, this script generates specialised PyMOL scripts by loading classes at runtime. The library has been used to create interfaces for Map and STAMP but it has been designed so that it can be easily extended to create similar interfaces for other fold recognition or structure comparison methods.

References

- (1) Russell, R. B., Copley, R. R. and Barton, G. J. (1996), "Protein fold recognition by mapping predicted secondary structures", *J. Mol. Biol.*, 259(3), 349-365
- (2) R. B. Russell & G. J. Barton, Multiple protein sequence alignment from tertiary structure comparison, *PROTEINS: Struct. Funct. Genet.*, 14, 309-323, 1992
- (3) DeLano, W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, San Carlos, CA, USA. <http://www.pymol.org>

Number: 65

Authors: Jan Reichert¹, Kristina Mehliß², Juergen Suehnel³

Institute of Molecular Biotechnology, Jena Centre for Bioinformatics, Jena / Germany, ¹jr@imb-jena.de

²mehliß@imb-jena.de ³jsuehnel@imb-jena.de

Title: The IMB Jena Image Library of Biological Macromolecules - Recent Developments



Short abstract: The database (<http://www.imb-jena.de/IMAGE.html>) is aimed at a better dissemination of information on three-dimensional biopolymer structures with an emphasis on visualization, classification and analysis. It provides basic information on the architecture of biological macromolecules and includes an Atlas of Macromolecule Structures with all PDB and NDB entries.

Full abstract: The IMB Jena Image Library of Biological Macromolecules (<http://www.imb-jena.de/IMAGE.html>) is aimed at a better dissemination of information on three-dimensional biopolymer structures with an emphasis on visualization and analysis [1]. A division on Basic Information on Biological Macromolecules includes, for example, the Amino Acid Repository, the Base Pair Directory, information on Nucleic Acids Nomenclature and Structure as well as on Structural Elements in Proteins. Also introductions to X-ray crystallography, NMR spectroscopy, Fourier Transform Infrared Spectroscopy and Circular Dichroism are available.

The Atlas of Macromolecule Structures provides access to all structure entries deposited at the Protein Data Bank (PDB) and Nucleic Acid Database (NDB). It offers many tools for an in-depth analysis of individual structures with an emphasis on visualization. Analysis tools are designed both for complete structures and for structure parts such as ligands, active sites, cis-peptide and SS bonds.

Given the increasing number of known three-dimensional biopolymer structures and with the data explosion in other fields of biology there is an urgent need for classification systems and for up-to-date and reliable cross-referencing schemes to other databases. To fulfill these needs the Image Library includes among other classification schemes the Hetero Components and Site Databases. A comprehensive bending classification of nucleic acid structures provides for currently more than 1150 entries tables listed by the extent of bending or by the shape of the helical axis. A genus/species classification of all PDB structures takes into account information from the NCBI taxonomy database. It includes a PDB structures species timeline that easily identifies new species entering the PDB and allows to search for scientific and common species names. The most recent additions to the Image Library features are a Gene Ontology (GO) interface, a user-friendly SCOP domain viewer, the platform-independent Java-Viewer Jmol and in-depth information on all currently known three-dimensional structures of peptaibols. The latter group of compounds includes also structures deposited at the Cambridge Structural Database that are thus not available from the PDB. The atlas pages include links to about 30 other databases or webtools, such as SWISS-PROT, InterPro, SMART, Pfam, ProtoMap, SYSTERS, WHAT-CHECK, Procheck and the Uppsala Electron Density Server. More recently disease-related information

resources such as OMIM have been included. Finally, there are ongoing efforts to link PDB structural information to genomics resources. Examples that are already operational are GeneCensus, PRESAGE and the human genome browser Ensembl.

In addition to this scientific usage, the Image Library is also used as an educational resource and has received attention from the public. Images from the database have been used in many books, journals, newspapers and exhibitions. For example, the journal RNA uses cover images from the Image Library for all 2003 and 2004 issues. The database has been featured as Site of the Month by *Science* in April 2001 and is also included in the listing of The Web's Best Sites of the Encyclopedia Britannica.

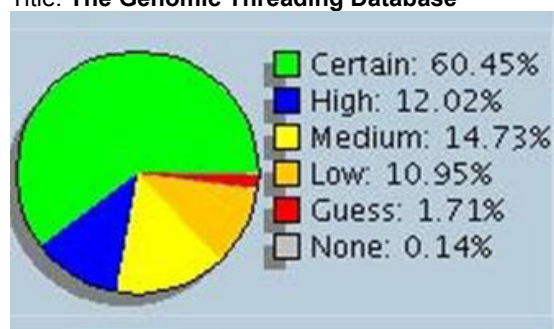
[1] Reichert, J., Sohnel, J. The IMB Jena Image Library of Biological Macromolecules - 2002 Update. *Nucleic Acids Res.* **2002**, 30, 253-254.

Number: 66

Authors: Liam J. McGuffin¹, Stefano A. Street², David T. Jones³, Soren-Aksel Sorensen

¹University College London, l.mcguiffin@cs.ucl.ac.uk ²s.street@cs.ucl.ac.uk ³dtj@cs.ucl.ac.uk

Title: The Genomic Threading Database



Short abstract: The Genomic Threading Database (GTD) currently contains structural assignments for the proteins encoded within the genomes of 174 organisms. Assignments are carried out using a version of GenTHREADER which is distributed using grid technology. The resulting predictions are deposited in a relational database accessible via a web interface at <http://bioinf.cs.ucl.ac.uk/GTD>.

Full abstract: Comprehensive and reliable annotation databases play an essential role in the interpretation and exploitation of the deluge of information resulting from genome sequencing projects. The Genomic Threading Database (GTD) is a new dedicated structural annotation database available on the web at <http://bioinf.cs.ucl.ac.uk/GTD>. A modified version of the genomic fold recognition method GenTHREADER (1,2) is the key part of the annotation system. This method enables reliable, sensitive and selective detection of remote homology between protein sequences and known folds. Grid technology is harnessed to speed up the process of accurately assigning structures to proteins from complete proteomes.

The GenTHREADER method consists of a feed-forward neural network which is trained to combine sequence alignment scores, length information, pairwise and solvation potentials derived from threading into a single score representing the likelihood of an evolutionary relationship between two proteins. The recently modified version also makes use of profiles seeded by structural alignments, bidirectional scoring and PSIPRED predicted secondary structure in order to maximize reliability and coverage (2).

The trade-off for more reliability and coverage is a reduction in the speed at which predictions are made. To counter this, annotation jobs are distributed across clusters of processors at University College London and Imperial College, London using grid technology.

The Globus toolkit—GT2 (<http://www.globus.org>)—is used to communicate between the remote sites. GT2 provides security between sites and secure job submission. In conjunction with the GT2, the jobs are scheduled using Sun Grid Engine (SGE, <http://www.sun.com/gridware>). The combination and use of these technologies results in a markedly improved throughput performance.

Using this system to distribute GenTHREADER jobs allowed us to rapidly predict the folds of proteins within the genomes of 174 organisms. Proteomic sequences were downloaded from a variety of sources; version numbers and the locations of the sequences are listed at <http://www.e-protein.org>. The resulting predictions and corresponding sequence alignments were uploaded into tables within a MySQL relational database (<http://www.mysql.com>).

A web interface was developed in order to allow all users to carry out keyword searches on the database (<http://bioinf.cs.ucl.ac.uk/GTD/>). GTD list files are also accessible which contain structural assignments for any given organism. An additional link is also provided to a summary statistics page, which enables users to access data on the coverage and reliability of annotations for each organism.

Password protected interfaces have also been developed in order to allow trusted users to automatically update and maintain the database through a web browser. These interfaces include; submission of new or updated proteomes to clusters for annotation and subsequent tracking of annotation progress, configuring the resulting

raw annotation data and uploading the data into the GTD and generating useful statistics concerning the uploaded data, for example, on coverage and reliability of annotations and fold frequencies per genome.

Since more accurate protein structure predictions can be gained from a consensus of methods (3) it is pertinent to combine the results from several annotation databases together. The e-Protein project (<http://www.e-protein.org>) is a pilot initiative which proposes to combine structural and functional annotation databases from University College London, Imperial College London and the European Bioinformatics Institute, through a single interface using the Distributed Annotation System (4). In addition, it is proposed that the workload of all annotation jobs will eventually be distributed across processing clusters at each site using a similar grid system to the GTD.

References

1. Jones, D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, 287, 797–815.
2. McGuffin, L.J. and Jones, D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, 19, 874–881.
3. Ginalski, K., Elofsson, A., Fischer, D. and Rychlewski, L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, 19, 1015–1018.
4. Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R. and Stein, L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, 2, 7.

Number: 67

Authors: Hannes Ponstingl¹, Janet M. Thornton²

European Bioinformatics Institute, EMBL-EBI, ¹hpo@ebi.ac.uk ²thornton@ebi.ac.uk

Title: Structure and Sequence Characteristics of Oligomeric Proteins



Short abstract: The three-dimensional crystal structures of oligomeric proteins are studied with respect to physico-chemical and geometric properties and sequence variation. Feature combinations are evaluated for the prediction of protein-protein interaction sites on the subunit surface.

Full abstract: Features of soluble, oligomeric proteins are reviewed to shed light on the formation of protein assemblies from a structural perspective. The features comprise physico-chemical and geometric entities as well as sequence conservation signals. They are compiled on low-redundancy sets of crystal structures of homomeric and heteromeric proteins of different subunit multiplicity. Crystal structures of proteins which are thought to function as monomers provide a control group.

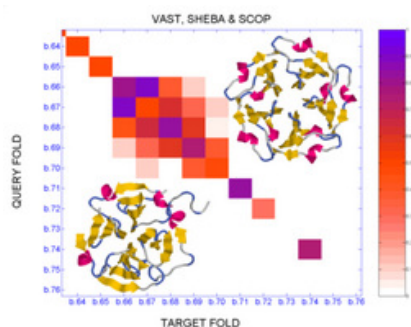
Interdependencies between various features are studied and informative feature combinations are evaluated for the prediction of interaction sites on the surface of a protomer using its structure alone or in combination with set of homologous sequences.

Number: 69

Authors: Vichetra Sam¹, Chin-Hsien (Emily) Tai², Peter J. Munson³, Jean Garnier, Jean-François Gibrat, Byungkook Lee

MSCL/DCB/CIT/NIH/DHHS, ¹vsam@helix.nih.gov ²taic@pop.nci.nih.gov ³munson@helix.nih.gov

Title: Comparison of Protein Structural Comparison Methods, VAST and SHEBA, with the SCOP Classification, using Statistical Methods



Short abstract: We conducted a comparison of the structural comparison methods VAST and SHEBA with the structural classification SCOP. Higher agreement with SCOP was obtained when combinations of scores from VAST and SHEBA were used. Fold-classification confusions were quantified. In depth visual analysis of the structures of confused folds will be presented.

Full abstract: Abstract

Many protein structural comparison/alignment methods have been developed over the year and they are extensively used for the classification of protein domains. It has been noted, however, that different methods give rise to significantly different results. These differences will affect the protein domain classification databases and our view of the protein structure universe.

In order to study the kind of differences that result from using different methods and to gain insight into the mechanism by which these differences arise, we conducted an in depth comparison of VAST, SHEBA and SCOP. SCOP is a manually curated protein domain structural classification database. It is often considered as the gold standard of protein structure classification. VAST is the prototype of several programs that achieve alignment by clustering vectors that represent secondary structural elements. It is the engine that NCBI uses in Entrez to find structural neighbors of a given protein. SHEBA is a rather unusual structure comparison program in that it uses sequence homology and structural profile for each residue to obtain the initial alignment.

The results of this work cover several aspects. First, we have developed an objective method based on statistics for comparing two structure comparison methods against a standard classification database like SCOP. Comparison of SHEBA and VAST using different scoring schemes and cutoff values has revealed their respective weaknesses and possible ways for future improvement. The comparison with SCOP using different cutoff values showed that both SHEBA and VAST give results that agree well with SCOP. For both methods, this agreement increases when two different score functions are combined. Combination of scores from the two methods brings higher agreement with SCOP than achieved by either method using a combination score separately for each method.

But important fold assignment confusions have also been noted: Some domain pairs that belong to the same SCOP fold are considered to be not structurally similar by SHEBA or VAST. Conversely, some domain pairs that are considered to be similar by SHEBA or VAST belong to two different SCOP folds. These expected discrepancies between automatic comparisons and a manual curation are then quantified for several scoring schemes and their combinations.

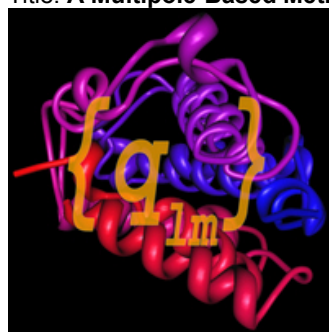
Results of in depth visual analysis of the structures of highly populated and confused folds will be presented. Preliminary indications are that many, but perhaps not all, discrepancies arise from the fact that human-based classification relies on additional criteria such as the function or other biological properties of the domain.

Number: 70

Authors: Apostol Gramada¹, Philip E. Bourne²

University of California San Diego, ¹agramada@sdsc.edu ²bourne@sdsc.edu

Title: A Multipole-Based Method for Comparison of Protein Structures



Short abstract: We present a method for the comparison of protein structures using a hierarchical set of shape descriptors. It is based on the use of a multipolar representation of the quantitative property of interest. Possible functions for protein similarity calculation are discussed and illustrated with test calculations.

Full abstract: Multipolar representations are largely used, for example, in the description of the interaction between a distribution of charge and an external electric potential. In that context, multipoles of a given rank couple with an inhomogeneity of the corresponding order of the potential field. From a different perspective, the set of multipolar quantities can be seen as a hierarchical set of descriptors for the shape of the distribution of charge and in general any scalar quantity which satisfy Laplace's equation. This opens the possibility for their use in the quantitative description of protein shape. In this sense, the approach has been used on a limited scale for the comparison of small molecules (1). Here we show how it can be extended to enable the comparison of protein structures. In particular, we describe different possible functions that can be used for the calculation of the similarity of a pair of proteins and illustrate the performance of this approach on some test cases.

References:

(1). B. D. Silverman, and Daniel E. Platt, Comparative Molecular Analysis (CoMMA): 3D-QSAR without Molecular Superposition, *J. Med. Chem.* 1996, **39**, 2129-2140.

Number: 71

Authors: Goran Neshich¹, Roberto Higa², Adauto Mancini, Paula Kuzer, Michel Yamagishi, Renato Fileto

¹EMBRAPA / CNPTIA, neshich@cnptia.embrapa.br ²Embrapa / CNPTIA, roberto@cnptia.embrapa.br

Title: Selecting a Set of the Protein Structure Descriptors Uniquely Defining the Active Site Residues



Short abstract: Is there a set of parameters (protein structure descriptors) which can define UNIQUELY an amino acid ensemble coinciding with the active site of a given protein? Our newly assembled STING_DB allowed us to test and confirm this hypothesis on several protein structures, including HIV-integrase.

Full abstract: Abstract:

We asked a very simple question: Is there a set of parameters (protein structure descriptors) which can define UNIQUELY an amino acid ensemble coinciding with the active site of a given protein? Our newly assembled, structure related, database: STING_DB (this database is, as far as we are aware, the largest collection of its kind), allowed us to test this hypothesis on several protein structures) and to confirm the existence of such set of parameters. Here we will present our in silico experiment done by Java Protein Dossier interface and database, on HIV-integrase, where we have obtained a very clear confirmation on the existence of such set of parameters.

Introduction:

Information technologies for heterogeneous database access, pattern discovery, and systems and molecular modelling have evolved to become core components of the today's molecular biologist's activities. As a consequence, we will continue to witness a shift in the balance between in silico modelling and experimental work. Rapid advances in the gene sequencing combined with volume increase of databases of protein sequences and structures have created a new environment which is far from being fully explored.

JavaProtein Dossier (JPD) is a new concept, database and visualization tool providing one of the largest collections of the physicochemical parameters describing proteins' structure, stability, function and interaction with other macromolecules. JPD (JavaProtein Dossier) communicates a large set of the physicochemical properties of proteins at a residue-by-residue level through a unique graphic interface. Furthermore, data grouping allows us to generate different parameters with the potential to provide new insights into the sequence-structure-function relationship.

Selecting parameters related to a protein function:

In JPD, residue selection can be performed according to multiple criteria. JavaProtein Dossier allows the user to make very informed decisions about the possible role of specific amino acids in defining the function of a protein. It also helps in deciphering what effect a specific mutation will possibly have on the structure and function of a protein, specifically by observing the changes in intra- and interface contact signatures, the sum of the energy values for established contacts, the electrostatic potential at the surface and the conservation. Selection of amino acids can be made with any combination of conditions, permitting powerful identification of functionally/structurally important regions or sites. For example, the user can select residues located at the interface between two chains, having a negative electrostatic potential at the surface they form, which are preserved in terms of evolution. (For detailed examples, see the JPD help manual: JPD Select Residues option).

HIV-1 integrase is an essential enzyme in the life cycle of the virus, responsible for catalyzing the insertion of the viral genome into the host cell chromosome; it provides an attractive target for antiviral drug design!!

The evidence was obtained from site-directed mutagenesis experiments in which it was demonstrated that even the most conservative substitutions of any of the three absolutely conserved carboxylate residues, D64, D116, and E152 (the so-called D,D-35-E motif), abolished catalytic activity!!!

By combining select ranges for following structure parameters:

1. Relevant site: Residue Location: Surface
2. Residue property: Residue type: Hydrophilic
3. Conservation: HSSP: Relative Entropy < 45
4. Conservation: HSSP: Relative Entropy 100 < 40
5. Conservation: HSSP: Evolutionary Pressure < 40
6. Conservation: HSSP: Reliability > 25%
7. Conservation: SH2Qs: Relative Entropy < 30
8. Conservation: SH2Qs: Relative Entropy 100 < 30
9. Conservation: SH2Qs: Evolutionary Pressure < 30
10. Conservation: SH2Qs: Reliability > 50%
11. Physical-Chemical: Electrostatic Potential: Average < -30 kT/J/mol
12. Geometric: Pocket/Cavity in Complex: Volume > 0

We have obtained an ensemble of amino acids which coincides with the critical three residues identified as active site of the HIV integrase: so-called D, D-35-E motif (see Fig 1.).

Conclusion:

In this presentation we are capitalizing over the JPD capability of being able to group amino acids into subsets which have a definite relationship to biological function. The application discussed below illustrates how powerful JPD can be when used in research and teaching.

Availability:

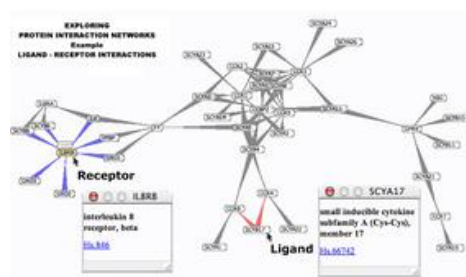
JPD is freely accessible (within the Gold Sting Suite) at <http://sms.cbi.cnptia.embrapa.br>, <http://mirrors.rcsb.org/SMS>, <http://trantor.bioc.columbia.edu/SMS> and <http://www.es.embnnet.org/SMS/> (Option: Java Protein Dossier).

Number: 73

Authors: **Javier De Las Rivas¹**, **Carlos Prieto**, **Alberto De Luis**

CIC, CSIC-USAL, ¹jrivas@usal.es

Title: **Exploring Protein Interaction Networks and Functional Relationships at a Genomic Scale with a Agile Browser**



Short abstract: Development of the bioinformatic tool APIN (Agile Protein Interaction Network browser) design to view and browse the nodes and links/edges of protein-protein interaction databases. The interface allows to navigate into complex interactome databases focusing on some areas, or on some specific proteins, by applying restrictions to the queried data.

Full abstract: As a first step to facilitate the understanding of protein-protein interaction networks we are developing a bioinformatic tool called APIN (Agile Protein Interaction Network browser). This tool will be designed to view and browse the nodes and links/edges of a protein interaction databases like DIP. The nodes are the proteins and the links or edges are the types of interaction. APIN is based on a JAVA applet called TouchGraph and displays the data stored in a MySQL database of interactions. The APIN browser reads data from the database and displays them in an interactive customizable way, like a browser. In this way, using an intuitive and clear interface one can navigate into complex interactome databases focusing on some areas, or on some specific proteins, by applying restrictions to the queried data. The tool is still in early development and we are implementing its capacity and power to browse over complex protein interaction networks, including three possible levels of association that we have defined: co-interacting proteins, defined as proteins that have physical interaction; co-related proteins, defined as proteins that are involved in the same biomolecular activity but that do not interact physically; co-located proteins, defined as proteins that work in the same cellular compartment (De

Las Rivas and De Luis, 2004).

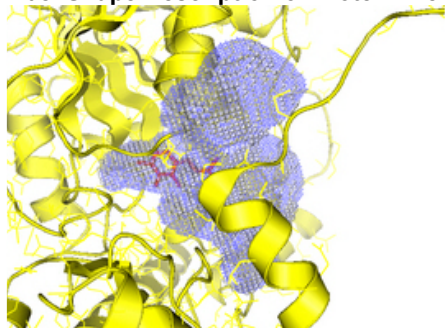
– De Las Rivas J. and De Luis A. (2004) Interactome data and databases: different types of protein interaction
Comparative and Functional Genomics 5: 173–178.

Number: 75

Authors: **Rafael Najmanovich¹, Richard J. Morris, Roman Laskowski, Janet M. Thornton**

¹European Bioinformatics Institute, rafael.najmanovich@ebi.ac.uk

Title: **Shape Description of Protein Clefts**



Short abstract: In the present work we describe the utilization of hybrid ellipsoids to model the shape of protein clefts. The small number of parameters describing the hybrid ellipsoid model can be used to assess binding site shape similarities. The method has applications in the prediction of protein function from structure.

Full abstract: Protein structures have been used to explain protein function for several years. However, the prediction of protein function from structure is still a problem in structural biology. In recent years, with the advent of structural genomics projects, this problem has become more relevant as the number of solved structures of proteins with unknown function grows. There are several methods exploiting different characteristics of the sequence and structure of the protein with unknown function for the prediction of function by similarity, i.e., by means of finding a matching protein whose function is already known. One characteristic not yet fully exploited in functional prediction methods is the shape of clefts in the surface of the protein. Similarities between a cleft and a known binding site may provide indications of the function and/or regulation of the unknown protein. In the present work we describe a method to model the shape of protein clefts using hybrid ellipsoids. Hybrid ellipsoids (Vaerman et. al., 1999) are parametric deformable models used for multidimensional object representation, a particular case of hybrid hyperquadratics. In three dimensions, an ellipsoid is completely determined by 9 parameters (center coordinates, length of half axes and rotation angles). A hybrid ellipsoid model can be seen as being composed of an ellipsoid (referred to as global) with local deformations introduced by means of weighted decreasing exponential ellipsoid terms that preserve the continuity of the surface. Thus, the overall hybrid ellipsoid model of a three dimensional object is described by a set of $9 + (1+9)N$ parameters, where N is the number of local ellipsoids with the extra parameter for each local ellipsoid being the weight factor.

Protein clefts are determined using the SURFNET algorithm (Laskowski, 1995). For a given cleft, a genetic algorithm optimization is used to locate the best placement of the global ellipsoid, i.e., the position that minimizes the number of points in a cubic grid that are either inside the global ellipsoid but outside the protein cleft or outside the global ellipsoid but inside the protein cleft. Once a global ellipsoid is chosen, local ellipsoids are introduced in sequence in a similar manner. The procedure is stopped once the overall improvement introduced by adding a local ellipsoid is below a threshold value or a maximum number of local ellipsoids have been added. Once a final hybrid ellipsoid model of the cleft is obtained, the overall shape of the cleft is described (within the chosen level of detail dictated by the number of local ellipsoids) by the set of $9+10N$ parameters. Fast comparisons of the overall shape of clefts can thus be made through the comparison of the hybrid ellipsoid parameters.

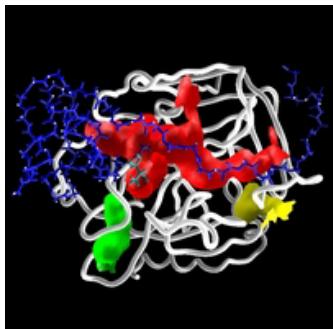
Hybrid ellipsoid models have the potential advantage of being able to accommodate differences between clefts (e.g., due to side chain movements) through differences in the placement and/or number of ellipsoids. One important characteristic of clefts other than the shape is the physico-chemical properties of the protein atoms in its surface. Such characteristics can be associated to the surface of the hybrid ellipsoid model and used as part of the overall comparison procedure to assign a similarity score that takes in consideration not only the shape of the cleft but its physico-chemical properties as well. This work describes the implementation of the method, presents several hybrid ellipsoid models of proteins clefts as well as their use and limitations in similarity comparisons.

Number: 78

Authors: **Murad Naya¹, Barry Honig²**

HHMI/Columbia University, ¹mn216@columbia.edu ²bh6@columbia.edu

Title: **Protein Surface Cavities and their Role in Protein-Protein Interactions**



Short abstract: We present a new method, DAMASCUS, Detection and Analysis of Molecular Surface Cavities and Ullage Space, for the identification of protein surface cavities. The method is used to analyze 3556 non-redundant protein-protein complexes. Various properties of surface cavities in and outside protein-protein interfaces are described and their functional implications discussed.

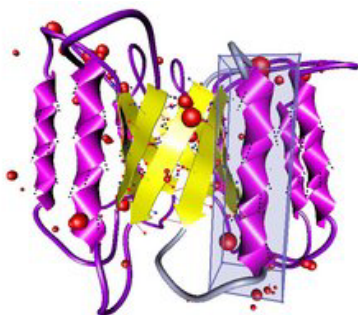
Full abstract: We present in this paper a new method, DAMASCUS, Detection and Analysis of Molecular Surface Cavities and Ullage Space, for the identification and the accurate characterization of protein surface cavities. The method is used to analyze molecular surface cavities in a comprehensive set of 3556 non-redundant protein-protein complexes comprised of 2216 homodimers and 1340 heterodimers. Three sub groups of the dataset were singled out for detailed analysis: enzyme-protein, antibody-antigen, and other immune system complexes. We show that protein surfaces typically have a small number of non-trivial cavities (an average of 8.98) and very few large ones (an average of 1.76 for surface cavities greater than 3003 Å³ in volume). We also show that surface cavities likely play an important role in molecular recognition. In particular, residues in surface cavities are more conserved than the rest of the protein surface, with cavities in the protein-protein interface somewhat more conserved than cavities outside the interface. Surface cavities are also structurally rigid compared to the rest of the protein surface as judged by the B factors of surface atoms as well as the conformational entropy of surface residues. This has important implications for molecular recognition as structural rigidity increases binding specificity by demanding strict shape complementarity of the bound ligand. It also enhances the binding affinity to compatible ligands since a rigid, pre-organized binding site loses less entropy upon binding than a flexible site. In addition, we find that shape complementarity is higher in protein-protein interfaces at surface cavities than it is in the rest of the interface. We find that surface cavities are not a consistent feature of protein-protein interfaces. However, case studies indicate that surface cavities, when they do exist at the binding site, often are functionally important. With the exception of enzyme-inhibitor complexes, only rarely (12.3%) does the largest surface cavity coincide with the protein-protein interaction site (more than 50% of cavity surface overlaps the binding site). However, in 41.6% of the cases the largest cavity has at least 1% overlap with the binding site.

Number: 79

Authors: Silvia N. Crivelli¹, James Lu², James Lu, Oliver Kreylos, Nelson Max, and Wes Bethel

Lawrence Berkeley National Laboratory, ¹SNCrivelli@lbl.gov ²TLu@lbl.gov

Title: ProteinShop: a Tool for Interactive Protein Manipulation



Short abstract: We describe ProteinShop, a program that creates and manipulates protein structures interactively. ProteinShop provides unique capabilities that enable its users to apply their accumulated biochemical knowledge and intuition during the interactive manipulation of protein structures.

Full abstract: In this laptop session we will demonstrate ProteinShop, an interactive visualization and manipulation tool designed to permit users to design tertiary structure of proteins using their knowledge and intuition. ProteinShop provides support for the following:

- Creation of proteins "from scratch" using the amino acid sequence and secondary structure specifications as input.
- Interactive manipulation of a 3D protein structure using a mouse-based user interface and an inverse kinematics method (commonly used in robotics) that changes dihedral angles along the backbone without

breaking the structure of the protein.

- Visualization of the distribution of backbone dihedral angles in selected regions of a protein over a pre-computed map of the Ramachandran energy calculated for each possible combination of (ϕ , ψ) dihedral angles.
- Interactive reassignment of secondary structure types to individual residues.
- Automatic β -sheet formation and ranking of these configurations using scoring and probabilistic functions.
- Interactive comparison and analysis of alternative structures through visualization of free energy computed during modeling.
- Accelerate discovery of low-energy configurations by applying local optimizations to user-selected protein structures.

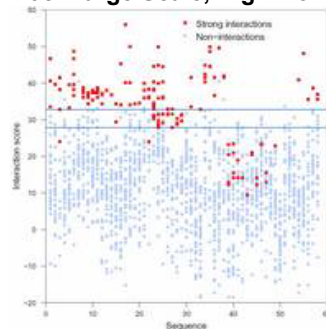
ProteinShop can be downloaded from <http://proteinshop.lbl.gov>.

Number: 80

Authors: **Jessica Fong**¹, **Amy Keating**², **Mona Singh**³

Princeton University, ¹jfong@cs.princeton.edu ²keating@mit.edu ³mona@cs.princeton.edu

Title: **Large Scale, High-Confidence Predictions of bZIP Protein-Protein Interactions**



Short abstract: We have developed a method for predicting protein-protein interactions mediated by coiled coils. Testing on interactions among nearly all human bZIP proteins, our method identifies 70% of strong interactions while maintaining that 92% of predictions are correct. Our work demonstrates a structural interaction motif for which large-scale, high-confidence predictions can be made.

Full abstract: A major challenge in structural bioinformatics is to develop methods for predicting protein-protein interactions at the genomic scale. The difficulty of computationally predicting protein structures suggests a strategy of concentrating first on interactions mediated by specific interfaces of known geometry. We approach this problem by focusing on a common and well-studied structural interaction motif, the parallel two-stranded coiled coil. We have developed a novel framework for predicting coiled-coil protein interactions that is able to use both genomic sequence data and experimental biophysical data.

Our method represents coiled coils in terms of their interhelical interactions and derives, from a base dataset of sequence and experimental data, a "weight" that indicates how favorable each residue-residue interaction is. We have shown that our method can make high-confidence predictions of interactions and non-interactions. In particular, testing on coiled-coil interactions among nearly all human and yeast bZIP transcription factors, our method identifies 70% of strong interactions while maintaining that 92% of predictions are correct. Additionally, using our scoring scheme, the highest scoring pairings for most sequences represent coiled-coil interactions. (See the representative figure, where each column shows the scores for pairings of one sequence that correspond to strong coiled-coil interactions and non-interactions.) Thus, a simple method for identifying the most likely partners for any sequence is to select the highest scoring pairings for that sequence.

Our method can be used to predict bZIP interactions in other genomes, as well as to predict coiled-coil interactions more generally. While whole- and cross-genomic approaches to predicting protein partners have had some success, our work is the first to demonstrate a structural interaction interface for which large-scale, high-confidence computational predictions can be made.

Number: 81

Authors: **Jessica C. Shapiro**¹, **Douglas L. Brutlag**²

Stanford University, ¹jessicas@stanford.edu ²brutlag@stanford.edu

Title: **Automatic Structural Motif Discovery using FOLDMINER**



Short abstract: FoldMiner, an unsupervised motif discovery algorithm, analyzes high scoring structural superpositions of a query protein with a database of targets and identifies structurally conserved regions of the query. These conserved regions then describe a structural motif shared among the query and its structural neighbors. Web access is provided at <http://foldminer.stanford.edu/>.

Full abstract: FoldMiner provides several features that are of particular use in the automatic analysis of protein structures. Its alignment scores are easily converted into distances between structures that are empirically found to obey the triangle inequality. Because LOCK 2 is a symmetric superposition algorithm, this distance is a metric that measures structural similarity between proteins. Pairwise alignments of protein structures that share a common fold are almost completely transitive at the level of secondary structure elements and are nearly transitive at the residue level. That is, given three protein structures, two of the three possible pairwise alignments predict the third alignment. These properties allow users to take advantage of information contained within pairwise structural alignments in order to detect similarities across multiple structures.

While a protein fold can be described by a motif that is common to all members of the fold, individual structures may have insertions consisting of entire secondary structure elements (SSEs) whose presence must be accounted for when assessing the statistical significance of structural similarities. We previously described an algorithm, FoldMiner, which addresses this issue by performing unsupervised motif discovery in the context of a structural similarity search. Given a query structure and a database of targets, FoldMiner automatically discovers the core fold shared among the query and its structural neighbors and then identifies additional targets that align to this core fold even when the query and/or the targets have additional secondary structure elements that may not align. Pairwise structural alignments are performed by LOCK 2.

FoldMiner operates by examining pairwise structural superpositions of the query structure to a database of targets in order to identify regions of the query that tend to be structural conserved among high scoring target structures. This information permits better discrimination between true and false positives by requiring that targets not only achieve high alignment scores, but also that they align to a specific, conserved region of the query. FoldMiner therefore favors alignments that are potentially local with respect to the query and target structures but that are global with respect to the query's core fold. The algorithm locates the core fold or motif shared among the query and its structural neighbors by identifying secondary structure elements (SSEs) whose positions are conserved among high scoring target structures. The scoring system then adapts to favor alignments to conserved SSEs, re-scores all alignments both to identify additional members of the fold and to remove false positives from the list of statistically significant hits, and iterates until the motif definition converges. This process is completely unsupervised and requires no prior knowledge of the size, location, or topology of the core fold or motif. The core fold as manifested in the query structure is described probabilistically in terms of the structural conservations of the query's SSEs and can also be viewed manually by depicting conserved SSEs in bright colors.

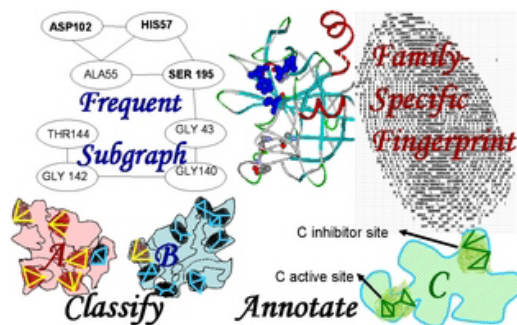
Figure Legend:

This figure shows four structural motifs discovered automatically by FOLDMINER. Secondary Structure elements colored in bright red are more highly conserved in the fold family than those which are darker. Beta-strands marked by arrows in the immunoglobulin-like fold are present in only a few family members.

Number: 82

Authors: Deepak Bandyopadhyay¹, Luke Huan², Wei Wang, Jack Snoeyink, Jan Prins, Alexander Tropsha
UNC Chapel Hill, ¹debug@cs.unc.edu ²huan@cs.unc.edu

Title: Graph Representations and Algorithms for Protein Family Classification and Functional Annotation



Short abstract: Protein family-specific fingerprints are residue packing patterns occurring frequently within a structural/functional family and rarely outside it. We compare three graph representations (contact distance, Delaunay and Almost-Delaunay) for finding fingerprints, discuss their biological significance, and use them for protein family classification and functional annotation.

Full abstract: Recurring substructural motifs in proteins reveal important information about protein structure and function. Motifs that are frequent among proteins in a structural/functional family and rare in the remaining PDB (the background) serve as fingerprints for the family, and often occur in active sites or other functional regions. Fingerprints represent residue packing patterns unique to a family, and cannot be determined from sequence comparison alone. We analyze the biological significance of some of the fingerprints we find, use them to build binary and multi-class classifiers, and finally describe techniques for functional annotation, eg. of structures from structural genomics projects.

For fast automatic motif extraction, we represent proteins as graphs, with nodes at residue C α s and edges joining residues consecutive in sequence or neighboring in space. Then we apply a new algorithm for mining maximal frequent subgraphs (Huan et al, SIGKDD2004) to find common subgraphs and derive family fingerprints. Using maximal frequent subgraphs is more robust than frequent induced subgraphs (Huan et al, RECOMB2004) under variations in local connectivity.

Subgraph mining is slow for dense graphs, such the contact distance (CD) graph where all residues within a distance threshold are considered neighbors. We explore two alternate representations: the Delaunay and almost-Delaunay edge graphs.

The Delaunay tessellation defines the nearest neighbors in a 3D point set as 4-tuples that satisfy an "empty sphere" property, and has been widely used in the analysis of protein structures. The graph of edges of Delaunay tetrahedra shorter than the distance threshold (denoted DT graph) is sparser than CD, and thus much faster to search, but is potentially unstable; small perturbations arising from imprecision in the input coordinates lead to large changes in the DT. Thus DT graphs for a family may miss many common subgraphs.

Our solution is to use the almost-Delaunay edges AD(ϵ), edges that can satisfy the Delaunay property for some movement of all points by at most ϵ . The parameter ϵ in AD measures the allowed perturbation in the input data. AD edges are described in our geometric framework for modeling nearest-neighbor relations in imprecise points (Bandyopadhyay and Snoeyink, 2004). For the subgraph mining application, AD gives a parameterized graph that interpolates in density between DT and CD, and can find motifs as good as those from CD in a fraction of the time.

The performance of graph representations for fingerprint identification depends on the sequence and structural variability within the family. In highly homologous families such as the Serine Proteases, with large common substructures, DT graphs are adequate for finding fingerprints, AD(0.5) graphs are better but slow, and computation on CD graphs does not terminate in a reasonable time. In structurally diverse protein families such as the Protein Kinases, DT graphs do not find any significant fingerprints while AD(0.5) and CD graphs do, and even CD graphs run in a reasonable time.

In the Serine protease family, we identified 19 fingerprints occurring in less than 1 percent of the background proteins, and one of the largest of these is 8 residues long and contains the Asp-His-Ser catalytic triad present in all Serine Proteases. The first four residues Asp-Ala-His-Ser of this motif superimpose to 0.5Å in the catalytic region, reflecting the convergent active site geometry of both eukaryotic and prokaryotic Serine Proteases. In the Kinase family, 6 fingerprints occurred in less than 1.5 percent of the background, including a linear Leu-Met-Ala-Leu-Ile-Lys running from functional residues in hydrophobic pockets to the catalytic Lysine. This fingerprint did not superimpose, showing that connectivity but not geometry is preserved across the functional regions of this structurally diverse family.

To perform binary or multi-class classification of proteins in different SCOP families, we collected a set of fingerprints for each family, sorted them by decreasing density, decreasing size and increasing background frequency, and picked the top k ($k=3$) features for each family. We then represented each protein as a k -element vector counting the occurrences of each feature, ran a Support Vector Machine, and found the highest accuracy of classification (95% and 93% for binary and three-class) for AD(0.5), and 90% for DT and CD.

Since fingerprints are highly specific to particular protein families, we can analyze the ones found in a query protein to distinguish between possible family assignments, and annotate the query structure with its functional regions. The annotation problem is equivalent to subgraph isomorphism, an NP-complete problem. To make it tractable, we build indices for each fingerprint and the query structure, containing some pre-calculated information about the graphs. Index comparison can rule out the occurrence of a fingerprint in a query without checking for subgraph isomorphism. We tried several indices, calculated at each residue: compositional adjacency (storing adjacent pairs of amino acids), path vectors (storing amino acids along paths of length 1-3 from each residue); and local neighborhood information (summarizing subgraphs within distance 1-3 of a residue by their amino acid count).

As an example of the specificity of our algorithm, on average 80% of the fingerprints for the Eukaryotic Serine Proteases (ESP) family found by our local neighborhood index in ESP and non-ESP queries, were confirmed by adjacency matching. No true matches were missed. Finally, the largest ESP fingerprints were found, apart from ESP, in a few prokaryotic and viral serine proteases, and also in 139 proteins from other families in a non-redundant dataset of 5000 proteins. We are investigating the significance of these matches, some of which are enzymes with similar function to ESP. Thus, our structure-based functional annotation may discover proteins that are remotely related in overall structure but have similar function.

References:

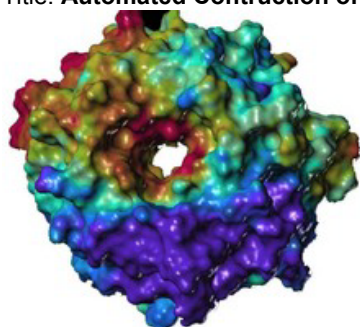
J.Huan, W.Wang, J.Prins and J.Yang. SPIN: Mining Maximal Frequent Subgraphs from Graph Databases. Proc. ACM SIGKDD 2004, to appear.
 J.Huan, W.Wang, D.Bandyopadhyay, J.Snoeyink, J.Prins and A.Tropsha. Mining Protein Family Specific Residue Packing Patterns From Protein Structure Graphs. Proc. RECOMB 2004, pp. 308-315. Invited to JCB Special Issue on RECOMB.
 D.Bandyopadhyay and J.Snoeyink. Almost-Delaunay Simplices: Nearest Neighbor Relations for Imprecise Points. Proc. Symposium on Discrete Algorithms (SODA) 2004, pp. 403-412.

Number: 83

Authors: Agnes Tan¹

¹Institute of Molecular and Cell Biology, mcbtanlc@imcb.a-star.edu.sg

Title: Automated Construction of WD40 proteins



Short abstract: We have written scripts in Sybyl Programming Language to automate construction of models of any beta propeller protein given a template and positions of beta strands, and also for the general case of modeling any protein by direct mutation (with optimization) of a suitable template based on a ClustalW alignment

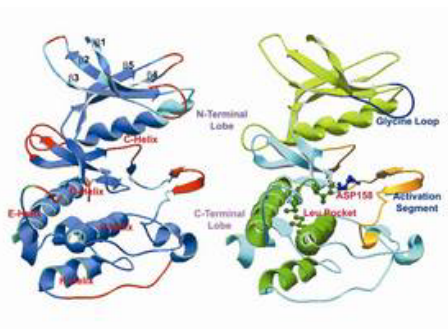
Full abstract: Scripting in molecular modeling is a powerful tool as it obviates the necessity of performing many tedious tasks manually. We have written scripts in Sybyl Programming Language to perform two distinct, though somewhat overlapping, tasks, namely, (i) to build beta propeller proteins from a template, and (ii) for the general case of modeling proteins via direct mutation of a template, based on a ClustalW alignment. For (i), the script reads as input file the number of blades in the propeller and the positions of the first amino acid in each beta strand (we have written a Perl script to generate such a file for a particular class of WD repeat proteins), along with the usual information. Aside from mutation, length of loops is a major issue. Rebuilding of loops appears to yield results of comparable quality to direct modeling of insertions and deletions. Optimization is achieved through molecular dynamics with dynamic definition of aggregates, as the program loops through the various beta strands. The algorithm for (ii) is based on the insertion/deletion model, and the length of the chain that is subject to optimization when the chain is modified is specified by the user. In all, there is no human intervention except for refinement of side chains at the end.

Number: 84

Authors: Alvin Ng¹

¹Institute of Molecular and Cell Biology, Singapore, ngyj@imcb.a-star.edu.sg

Title: Homology Model and Substrate Recognition of Yeast Prk1p Kinase



Short abstract: Prk1p is a serine/threonine kinase involved in actin cytoskeleton organization in yeast *Saccharomyces cerevisiae*. A homology model was created to further understand the substrate recognition of the **LxxQxTG** motif.

Full abstract: Prk1p regulates the actin cytoskeleton by modulating the activity of the cytoskeleton-related protein Pan1p (Zeng and Cai, 1999). Pan1p is necessary for normal organization of the actin cytoskeleton during budding. Prk1p phosphorylates Pan1p in vitro on threonine residues located in LR (Long Repeats) domain of Pan1p with the repeating common motif LxxQxTG. The P-5 residue (5 residues before the phosphorylatable threonine in the direction of the N-terminal) of the substrate is observed experimentally to decrease in phosphorylation intensity when mutated from Leucine to Isoleucine to Valine to Methionine respectively (Huang et al., 2003).

Prk1p is an 810 residue long serine/threonine kinase with significant homology to other kinases in the kinase domain. However, there are no crystallized structures of Prk1p or any other members of its family. Multiple alignments and phylogeny of Prk1p showed that it belongs to a distinct family of kinases known as the ARK family.

An initial model of Prk1p was generated through the on-line automated comparative protein modeling server, SWISSMODEL (Guex and Peitsch, 1997). Approximately 70 templates were used to generate a model from residues 23 to 298 of the kinase domain of Prk1p. The Prk1p model has structurally conserved domains common to other documented kinases.

Figure: Ribbon model of Prk1p generated from PDBViewer. The model shows that Prk1p has kinase conserved structures. The catalytic residue D158 which is essential for kinase activity is labeled. The hypothetical P-5 pocket (Leucine pocket) is shown together with the 3 amino acids which make up the pocket.

The LxxQxTG motif is proposed to bind to the C-terminal lobe, similar to substrates of other documented serine/threonine kinases (Madhusudan et al., 1994). The phosphorylatable residue, Threonine, should have its hydroxyl group in proximity to the carboxyl side chain of the Prk1p catalytic residue D158 to facilitate phosphotransfer (Johnson et al., 1998).

Conserved residue K (K160 in Prk1p) on the kinase catalytic loop anchors the substrate to the C-terminal lobe. Another conserved residue T (T205 in Prk1p) located on the activation segment, H-bonds to an α -carbonyl of the substrate through its side chain hydroxyl group (Madhusudan et al., 1994) (See figure 4).

These interactions "anchor" the substrate and position the P-5 residue into a hydrophobic pocket made up of 3 hydrophobic residues (I117, M120 and I238 on Prk1p). SiteID from Sybyl6.8 (Tripos Inc, St Louis) was used to identify this pocket.

Molecular Dynamics and conformational search was performed on each P-5 mutant to determine their equilibrium docked conformations using Sybyl6.8. It is observed that I117 on Prk1p was important for hydrophobic interactions with the [LIVM]xxQxTG mutants.

The I117A mutant showed diminished phosphorylation, supporting the hypothesis that I117 pocket was necessary for P-5 substrate interaction. Molecular modeling was used to understand and predict how the Prk1p kinase phosphorylates its substrates.

Sybyl6.8 runs on an SGI Octane2 with 2 x R12000 400 Mhz Processors with a main memory of 1 Gigabyte. SWISSMODEL and SWISS PDB Viewer is available at <http://www.expasy.org/swissmod/SWISS-MODEL.html>. SWISS PDBViewer runs a 1.6GHz Windows Desktop machine.

Reference List

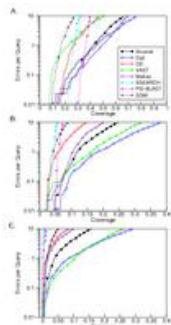
- Guex,N., Peitsch,M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18, 2714-2723.
- Huang,B., Zeng,G., Ng,A.Y., Cai,M. (2003). Identification of novel recognition motifs and regulatory targets for the yeast actin-regulating kinase Prk1p. *Mol Biol Cell* 14, 4871-4884.
- Johnson,L.N., Lowe,E.D., Noble,M.E., Owen,D.J. (1998). The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. *FEBS Lett.* 430, 1-11.
- Madhusudan, Trafny,E.A., Xuong,N.H., Adams,J.A., Ten Eyck,L.F., Taylor,S.S., Sowadski,J.M. (1994). cAMP-dependent protein kinase: crystallographic insights into substrate recognition and phosphotransfer. *Protein Sci.* 3, 176-187.
- Zeng,G., Cai,M. (1999). Regulation of the actin cytoskeleton organization in yeast by a novel serine/threonine kinase Prk1p. *J.Cell Biol.* 144, 71-82.
- Zeng,G., Yu,X., Cai,M. (2001). Regulation of yeast actin cytoskeleton-regulatory complex Pan1p/Sla1p/End3p by serine/threonine kinase Prk1p. *Mol.Biol.Cell* 12, 3759-3772.

Number: 85

Authors: Michael L. Sierk¹, William R. Pearson²

University of Virginia, ¹mls5w@virginia.edu ²wrp@virginia.edu

Title: Benchmarking Protein Structure Alignment Statistics Against the CATH Database



Short abstract: We present a large-scale analysis of several structure alignment methods and ability to detect homologs in the CATH domain database. The programs vary considerably in their performance, and the statistical estimates reported by the programs overstate the significance matches by orders of magnitude compared to the actual distribution of errors.

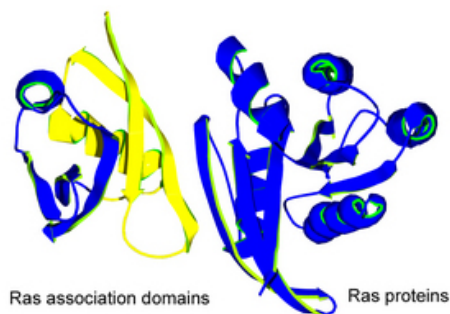
Full abstract: As more structures of proteins with unknown function are produced by high-throughput structure determination projects, it will become increasingly important to be able to determine whether a new structure is related to known structures in order to make inferences about function. There are a variety of alignment algorithms and similarity scores available for comparing two structures, but interpretation of such scores has traditionally been highly subjective. Here we present a large-scale analysis of different methods (Dali, Strucal, CE, VAST, Matras) against a common standard (the CATH database). We find significant variation in the ability of the tested programs to detect CATH homologs and topologs (domains with the same fold), and we find that the statistical estimates reported by the programs overstate the significance of a match by orders of magnitude compared to the actual distribution of errors. We also assess the errors made by the different programs, including whether the alignment or the scoring system is more important in discriminating true and false positives, and suggest ways to improve the performance and reliability of structure database searches.

Number: 86

Authors: Christina Kiel¹, Luis Serrano², Alfred Wittinghofer, Sabine Wohlgemuth

EMBL Heidelberg, Germany, ¹kiel@embl.de ²serrano@embl.de

Title: Recognizing and Defining true Ras Association (RA) Domains based on Protein Sequence and Structure



Short abstract: We are using homology modelling, and complex stability calculations using Fold-X in order to predict the affinities of Ras association (RA) domains in complex with Ras proteins. Calculated stabilities are compared with experimental data from Isothermal Titration Calorimetry experiments.

Full abstract: The members of the Ras-related super-family of GTP-binding proteins are small 20-25 kDa proteins, which cycle between an inactive GDP-bound and an active GTP-bound state. In the GTP-bound state, Ras proteins can interact with effector molecules as downstream targets thereby communicating signals into different pathways [1, 2]. Although the sequence homology is very low all Ras binding domains (RBDs) of effector proteins found so far show an ubiquitin like fold. Apart from these classical RBDs the RA (Ras association) domain family share some sequence homology with RBDs and are also predicted to have ubiquitin topology as well, but it is not known whether they really all bind to Ras proteins [3]. In this work we aimed at predicting the affinities of various RA domains in complex with Ras proteins based on the protein sequence using theoretical tools.

Using homology modelling (WHAT IF web interface) [4] and the protein engineering algorithm PERLA [5, 6] we build complexes of various RA domains in complex with Ras proteins and calculated the free energies of complex

formation using Fold-X [7]. Comparing these energy values with those derived from measurements using Isothermal Titration Calorimetry (ITC) shows a good correlation of experimental and calculated complex affinities. In addition, using mutant complexes we could also predict the critical hot-spot residues (Lys or Arg in $\alpha 1$, $\alpha 2$ and $\alpha 1$ of the RA domain) and residues important for specificity of RA domains in complex with Ras and Rap, confirming the quality of our modelled complexes.

Considering the large number of protein-protein interaction domains and variants thereof in many genomes it is highly desirable to develop theoretical tools for predicting which proteins can interact, before starting to investigate the physiological relevance of the interaction. In the future we aim at prediction of putative Ras-effector interactions using in silico screening at a more genome wide level.

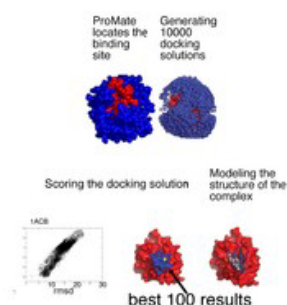
1. Bourne, H. R., Sanders, D. A. & McCormick, F. (1990). The GTPase superfamily: a conserved switch for diverse cell functions. *Nature*, 348, 125-132.
2. Bourne, H. R., Sanders, D. A. & McCormick, F. (1991). The GTPase superfamily: conserved structure and molecular mechanism. *Nature*, 349, 117-127.
3. Ponting, C. P. & Benjamin, D. R. (1996). A novel family of ras-binding domains. *Trends Biochem. Sci.* 21, 422-425.
4. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, 8, 52-56
5. Fissinger, S., Serrano, L. & Lacroix, E. (2001). Computational estimation of specific side chain interaction energies in α -helices. *Protein Sci.* 10, 809-818.
6. Lopez de la Paz, M., Lacroix, E., Ramirez-Alvarado, M. & Serrano, L. (2001). Computer-aided design of α -sheet peptides. *J. Mol. Biol.* 312, 229-246.
7. Guerois, F., Nielson J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* 320, 369-387

Number: 87

Authors: Kay Gottschalk, Hani Neuvirth, Gideon Schreiber¹

¹Weizmann Institute of Science, bcges@weizmann.ac.il

Title: Docking Protein Complexes from their Predicted Binding Site



Short abstract: We introduce a docking method which discriminates between the few correct and many wrong results by measuring the tightness of fit of two docked proteins at their predicted binding interface. The location of the binding interface is identified using ProMate (<http://biportal.weizmann.ac.il/promate>), a computer algorithm developed by us for this purpose.

Full abstract: Is the whole protein surface available for interaction with other proteins, or are specific sites pre-assigned according to their biophysical and structural character? And if so, is it possible to predict the location of the binding site from the surface properties?

These questions are quantitatively answered by probing the surfaces of proteins using spheres of radius of 10 Å on a database, of 57 unique, non-homologous proteins involved in heteromeric, transient protein-protein interactions for which the structures of both the unbound and bound states were determined. In structural terms, we found the binding site to have a preference for β -sheets and for relatively long non structured chains, but not for α helices. Chemically, aromatic side-chains show a clear preference for binding sites. While the hydrophobic and polar content of the interface is similar to the rest of the surface, hydrophobic and polar residues tend to cluster in interfaces. In the crystal, the binding site has more bound water molecules surrounding it, and a lower B-factor already in the unbound protein. The same biophysical properties were found to hold for the unbound and bound databases.

All the significant interface properties were combined into ProMate - an interface prediction program (<http://biportal.weizmann.ac.il/promate>). This was followed by an optimization step to choose the best combination of properties, as many of them are correlated. During optimization and prediction, the tested proteins were not used for data collection, to avoid over-fitting. The prediction algorithm is fully automated, and is used to predict the location of potential binding sites on unbound proteins with known structures. The algorithm is able to successfully predict the location of the interface for about 70% of the proteins. The success rate of the predictor was equal whether applied on the unbound database or on the disjoint bound database. A prediction is assumed to be correct if over half of the predicted continuous interface patch is indeed interface. Using ProMate to predict the location of the binding site, we went on to use this information for docking of protein complexes. Docking

algorithms produce many possible structures of a protein-protein complex, some of them resemble the correct structure within a RMSD of less than 3 Å. A major challenge in the field of docking is to extract the correct structure out of this pool, the so called “scoring”. Here, we introduce a new scoring function, which discriminates between the many wrong and few true conformations. The scoring function is based on measuring the tightness of fit of the two docked proteins at a predicted binding interface. The location of the binding interface is identified using ProMate. The new scoring function does not rely on energy considerations. It is therefore tolerant to low-resolution descriptions of the interface. A linear relation between the score and the RMSD relative to the “true structure” is found in most of the cases evaluated. The function was tested on the docking results of 21 complexes in their unbound form. It was found to be successful in 77% of the examined cases, defining success as scoring a “true” result with a P value of better than 0.1. The uniqueness of this approach allows us to dock also models of proteins, as a high-resolution structure is not required for interface predictions using ProMate.

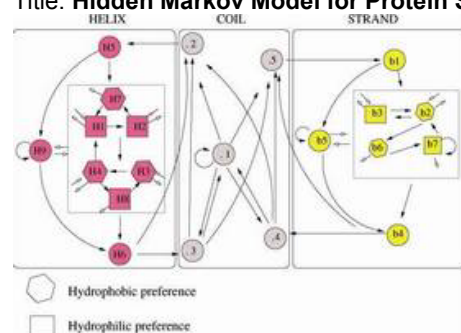
1. Neuvirth, H., Raz, R. & Schreiber, G. (2004). ProMate: A Structure Based Prediction Program to Identify the Location of Protein-Protein Binding Sites. *J Mol Biol* 338, 181-199.
2. Gottschalk, K. E., Neuvirth, H. & Schreiber, G. (2004). A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Eng Des Sel* 17, 183-189.

Number: 88

Authors: Juliette Martin¹, Jean-François Gibrat², François Rodolphe

¹INRA, jumartin@jouy.inra.fr ²Laboratoire Mathématique Informatique et Génome, INRA de Jouy-en-Josas, France, gibrat@jouy.inra.fr

Title: Hidden Markov Model for Protein Secondary Structure



Short abstract: We present an attempt to predict the secondary structure of proteins with Hidden Markov Models. We designed a 21-states model in which regular secondary structures are described by motifs. With single-sequence information the rate of good prediction of this model is about 65%, and 71% with profile information.

Full abstract: We are currently planning to develop a *de novo* method for predicting the 3D structure of proteins from their sequences. To begin with, we work on **local structure prediction**. As a first approach, we consider the widely-used three states description of proteins as alpha-helix, beta-strand or coil regions, before including a more accurate description of coils.

We use **Hidden Markov Models (HMMs)** to represent a protein sequence. The hidden process, we want to recover, is the secondary structure. The observed process is the amino acid sequence.

A non-redundant subset of 2530 protein structures (ASTRAL 1.65) is used to train our models and predict the secondary structure in a 4-fold cross validation procedure. Secondary structure is defined with our software program (unpublished) based on Phi-Psi angles and Calpha-Calpha distances. Parameter estimation is performed with maximum likelihood estimates, or with the EM algorithm, depending if the hidden sequence is known or not. Prediction is done using the posterior probabilities of the forward-backward algorithm.

We first test some basic 3-states models, with different memory schemes. The hidden process is always a M1 process. Over-fitting on the training set appears at low orders when increasing the memory of the observed chain. The best performance obtained is 60.3% of correct prediction, with a sparse-M1M2 model. In this model, observation at position *t* only depends on the observation at position *t-2*.

We then explore **more complex models** to take account of the main features of protein sequences. The best model we build is a **21-states M1M0 model** with 450 independent parameters. Helices (9 states) and strands (7 states) are modeled by structured patterns which can freely alternate with background distribution. First and last positions of segments are explicitly modeled. The coil description only distinguishes pre-helix/strand positions and post-helix/strand positions from the background distribution.

We incorporate the cyclic pattern for the amphipathic helix, the hydrophobic/hydrophilic pattern for beta-strands, and the most frequent and exceptional patterns we found in strands. These last patterns are identified using

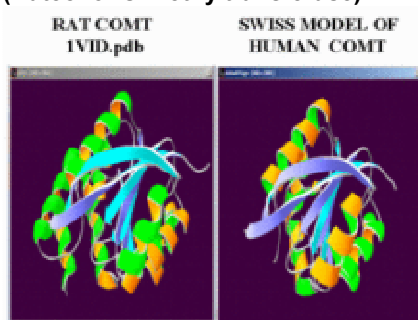
RMES which compares a word frequency with the expected frequency under a Markov model of given order. Q3 score of this model is 65.3%, and **66.3%** if we use the STRIDE assignment. We then used profile information. Predictions were made for related sequences detected by PSI-BLAST, and combined with Henikoff weights. Best Q3 obtained is **71.2%**.

Number: 89

Authors: C. Jayaprabha, N. Santhi¹, J. Kamesh², P. Chitra, N. Bharathi, S. Krishnaveni, S. Valarmathi, S. Sujatha and S. Seethalakshmi

¹Dept. of Bioinformatics, Sri Ramakrishna College of Arts & Science for Women, INDIA, santhigowri@rediffmail.com ²Sai's Biosciences Research Institute Pvt.Ltd., INDIA, jkbioinfo@graffiti.net

Title: *In Silico* Transfer of Neurotransmitter Transporter Motif Between Structurally Analogous Protein (Catechol-O-methyltransferase)



Short abstract: A homology model for Catechol-O-methyltransferase (COMT) of human was created, enabling us to analyse the active site. To analyse the function, the motif region is transferred from human COMT to rat COMT. The structure of the mutated rat sequence was compared with original rat COMT structure.

Full abstract: *In silico* methods can be used to design protein, based on stability and functionality using computational methods rather than laboratory procedures (Comet et al., 2000). The changes taking place due to the transfer of motif region from human Catechol-O-methyltransferase to rat Catechol-O-methyltransferase can be studied effectively using computational methods. This will provide insight for further development of the study about the function of neurotransmitter region of catechol-O-methyltransferase and its involvement in the Parkinson's disease.

Catechol-O-methyltransferase (COMT) catalyzes the transfer of the active methyl group from S-adenosyl-L-methionine to the two ring hydroxyl groups of catechols (Garner et al., 1978). The main function of COMT involves the elimination of biologically active or toxic catechols and their metabolites. COMT may modulate the neurotransmitter function of dopamine and norepinephrine in various mental processes through altering the rate of their metabolic inactivation in different parts of the brain.

COMT is found both in membrane bound form in post synaptic neurons and in cytoplasmic soluble form in the extraneuronal tissue like glial cells.

The sequence of the Human COMT was retrieved from Human Protein Reference Database (HPRD). The BLAST and ClustalW was performed to confirm the homologous sequence for human COMT as rat COMT seq. The motif sequence was identified from the Blocks searcher. PDB BLAST confirmed rat COMT as the structural homology for modeling the structure of human COMT using SWISSMODEL. The human sequence contained VLLELGAYCGYSVRMA in the neurotransmitter motif sequence. But some organism and rat possess neurotransmitter sequence with the change in the two amino acid val and leu (LVLELGAYCGYSVRMA)

An initial model of COMT of human was generated through the on-line automated comparative protein modeling server, SWISSMODEL (Guex, N., Peitsch, M.C. 1997). and was verified to evaluate the quality of the structure.

SAVS, a Structure Analysis and Verification Server was used to study the quality of the model structure generated by SWISSMODEL (Van G.W. 1996). The overall quality factor of the model was 96.429* and the 98.31* of the residues had an average 3D-ID score >0.2.

The *in silico* transfer was done with the help of SwissPDB viewer mutation tool. The structure of the mutated rat sequence was compared with original rat COMT structure. This may give us information about changes involved after the mutation and these changes may in turn affect the functions of the animal.

The results from both ExPASy tools and Predict Protein states that there is very less difference in the % composition of the secondary structures of both the sequences. These results infer that the mutated rat with the human motif sequences may also possess the same functionality and stability as that of the original rat COMT sequence.

References:

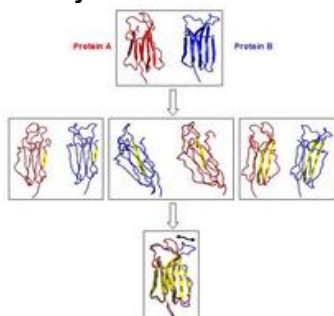
Combet C., Blanchet C., Geourjon C. and Del'age G., 2000 March, TIBS Vol. 25, No 3 [291]:147-150
 Garnier J, Osguthorpe DJ, Robson B., 1978, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol. Vol. 120, pp. no 97-120.
 Guex,N., Peitsch,M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714-2723.
 Van Gunsteren, W., 1996, Biomolecular Simulations: The GROMOS96 Manual and User Guide. VdF Hochschulverlag ETHZ.

Number: 90

Authors: **Giacomo De Mori¹, Raul Mendez², Didier Croes**

¹SCMBB-ULB (Belgium), giacomo@scmbb.ulb.ac.be ²raul@scmbb.ulb.ac.be

Title: **Analysis of Conformational Changes on Protein - Protein Complexes.**



Short abstract: An exhaustive analysis of protein - protein conformational changes is presented on protein complex structures deposited at the PDB. Secondary structure elements changing upon association are identified using a non-standard structural alignment method. A detailed description on the different conformational changes over the different protein complexed families is provided.

Full abstract: Intermolecular interactions, particularly those between proteins, have become a central focus in postgenomic biology. Tens of thousands of gene products are known or suspected to interact with many others, from genetic, biochemical, or bioinformatics studies forming millions of putative complexes. However, computational methods for protein - protein prediction face the problem of the structural flexibility of any of the components upon association.

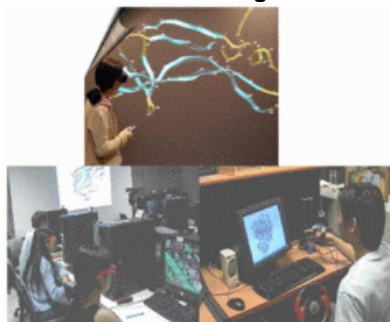
A highthroughput analysis of conformational changes over all known 3-D crystallographic structures available at the Protein Data Bank is presented here. In a first step, protein - protein complexes have been identified using SQL queries on a Oracle database system. Unbound protein components for each complex are then selected using Blast sequence alignment tool. Pairwise structural alignments have been performed to optimally superimpose unbound components to their corresponding bound structure using an in-house multiple linkage clustering algorithm. The aim of this last step is to search for protein regions (group of secondary structure elements) that remain invariable upon association in order to better understand the conformational changes. A final exhaustive analysis of the secondary structure elements involved on conformational changes is done, so a characterization of the different kind of structural movements among the different families of associated protein is proposed.

Number: 91

Authors: **YY Cai¹, BF Lu², ZW Fan, and KT Lim**

¹Nanyang Technological University, myycal@ntu.edu.sg ²mbflu@ntu.edu.sg

Title: **3D Structure Image Generator using Virtual Reality Technology for Protein Immersive Visualization**



Short abstract: We present our low-cost VR technology for interactive visualization and modeling of 3D protein structure. Emphasis will be placed on the affordable solution which integrates commodity hardware available and

specialized software we developed. In particular, visual and haptic sensorial channels for immersive interaction with the protein structure will be discussed.

Full abstract: As an enabling technology, Virtual Reality (VR) has been made significant progress recently and many successful applications can be found in areas of Science, Engineering, Education and Entertainment. With the milestone development in Human Genome Project, Life Science has entered a post-genome era which is more structure and function oriented. To better understand the insight of complex 3D molecular system, VR will be playing a special role augmenting the human sensorial capability.

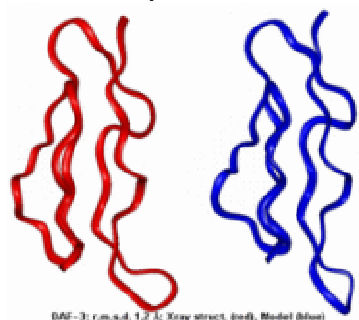
However, traditional VR is often referred as high-tech toys due to their huge initial investment. This also limits VR's popular use in schools, and laboratories. In this laptop session, we show how to use our low cost VR technology for interactive visualization and modeling of 3D protein structure. Emphasis will be placed on the affordable solution which integrates commodity hardware available and specialized software we developed. In particular, visual and haptic sensorial channels for the immersive interaction with the protein structure will be discussed.

Number: 92

Authors: Dinesh C. Soares¹, Paul N. Barlow², Dietlind L. Gerloff

¹Biocomputing Research Unit, University of Edinburgh, D.C.Soares@sms.ed.ac.uk ²Biomolecular NMR Unit, University of Edinburgh, Paul.Barlow@ed.ac.uk

Title: The CCP-Module Model Database - Automated Large-Scale Protein Structure Modelling of Individual Human Complement Control Protein (CCP)-Modules



Short abstract: A large-scale comparative protein structure modelling procedure was automated and applied to 136 representative individual CCP module sequences, including members of the regulators of complement activation family. Systematic comparisons of surface electrostatics provide new functional insight with respect to potential protein-protein interactions. All models are publicly available at:

<http://www.bru.ed.ac.uk/~dinesh/ccp-db.html>

Full abstract: The proteins belonging to the human regulators of complement activation (RCA) family - factor H (fH), complement receptor type 1 (CR1), membrane cofactor protein (MCP), decay accelerating factor (DAF), and C4b-binding protein (C4BP) - are crucial for ensuring that a complement-mediated immune response is proportionate and directed against the infectious agent. They are built up from homologous domains called complement control protein (CCP)-modules (also known as sushi domains and SCRs). Each module has between 55 to 71 amino acids and is separated from its neighbour by short linking sequences.

In light of the great medical interest in these proteins and the importance to provide as highly resolved 3-D structural information as possible, an automated large-scale protein structure modelling procedure was implemented for a representative set of 203 CCP module sequences taken from the SMART database [1]. The 3D structures of a diverse range of CCP modules, including twelve RCA modules, have been experimentally determined over the last decade. Their compact all-beta domain fold is characterised by a consensus "scaffold" of conserved buried residues, involving two disulphide bridges and a buried tryptophan. The set of target sequences were initially assigned to nine clusters on the basis of sequence similarity using an implementation of a hierarchical cluster assignment method [2] and extended through subsequent sequence comparisons using Hidden Markov models. Based on multiple sequence alignments for each cluster, comparative models were generated with the program Modeller [3] using the most similar homologues with known structures as modelling templates. Finally, an evaluation step in the procedure automatically confirms valid stereochemistry [4].

A total of 136 CCP module model structures from 36 different proteins have been produced in the first instance by this automated protocol. The large scale modelling process can easily be extended to include a larger representative set of human and non-human CCP module containing proteins. All models are made available publicly through the CCP-Module Model Database website <http://www.bru.ed.ac.uk/~dinesh/ccp-db.html>, which will be updated as more experimental template structures are solved and novel CCP-module sequences detected. A test comparison, of the individual CCP modules from the recently solved crystal structure of DAF (5) versus the models built using the automated modelling procedure, reveal high structural similarity (all equivalent

Calpha r.m.s.d. comparison for DAF~1: 1.6_Å, DAF~2: 2.2_Å, DAF~3: 1.2_Å and DAF~4: 1.9_Å), consistent with a successful modelling strategy.

The models provide a basis for rationalising the extensive mutagenesis and comparative functional data that is available for RCA and other CCP-module containing proteins. Furthermore, they allow large-scale surface electrostatic analyses [6,7,8] aiming to shed further light on the function of individual modules. Our first comparisons of sequence-based and surface-based clustering of the modules of Complement Receptors 1 and 2 (44 modules in total) identify modules whose surfaces may have undergone adaptive evolutionary change reflecting their involvement in specific protein-protein interactions and suggest new methodology for function prediction.

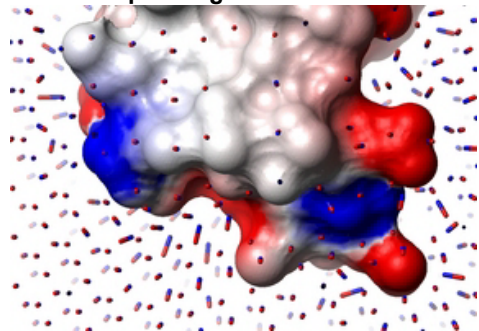
References:

- [1] SMART - Schultz et al. PNAS 95(11):5857-64 (1998).
- [2] Corpet F., Nucl. Acids Res., 16:10881-10890 (1988).
- [3] Modeller - Sali and Blundell, J. Mol. Biol. 234:779-815 (1993).
- [4] Procheck - Laskowski et al. J. App. Cryst. 26:283 (1993).
- [5] Lukacik et al. PNAS 101(5):1279-1284 (2004).
- [6] GRASP - Nicholls et al. Proteins: Struct., Funct., Genet. 11:281-296 (1991).
- [7] PIPSA - Blomberg et al. Proteins: Struct., Funct., Genet. 37:379-387 (1999).
- [8] NMRCLUST - Kelley et al. Protein Eng. 9:1063-1065 (1996).

Number: 93

Authors: Joaquim Mendes¹, Luis Serrano²
EMBL, ¹mendes@embl.de ²serrano@embl.de

Title: Incorporating an Accurate Solvation Model into Computational Protein Design



Short abstract: The Langevin Dipoles solvation model is one of the most accurate quantitative solvation models currently available. Here we describe its incorporation into one of the most successful methods for computational protein design (CPD): the Mean Field method. This work represents a major step forward toward first principle CPD methods.

Full abstract: In their natural environment, proteins are totally or partially immersed in an aqueous medium. As such, to predict their properties using Statistical Mechanics (SM) based methods, it does not suffice to consider the intramolecular interactions of the polypeptide chain. One must also consider the intermolecular interactions between the polypeptide chain and the water molecules of the aqueous medium, as well as the alterations in the interactions of the water molecules among themselves induced by the presence of the protein. These interactions contain both enthalpic and entropic contributions. For a protein-solvent system at thermodynamic equilibrium, they are quantified by the solvation free energy of the protein molecule.


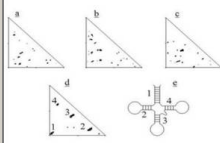
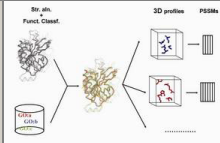
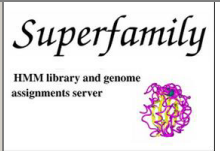


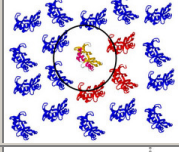
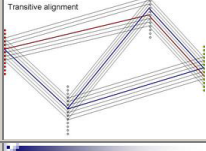
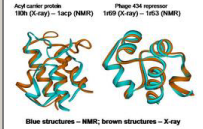
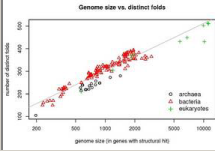
Several solvation models have been developed to estimate the solvation free energy of a given molecule[1-4]. In almost all methodologies for computational protein design (CPD), solvation free energy has been modeled using the Atomic Solvation Parameters (ASP) model[5] in combination with a distance dependent dielectric constant. Although this model can provide qualitative insights for simple problems, it is far too crude for the quantitative free energy estimation required for the complex problem these methodologies attempt to address[6]. The inaccurate modeling of solvation is probably one of the main causes of the limited success of current CPD methodologies[7].

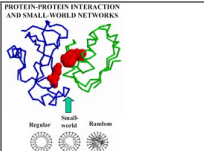
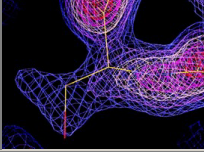
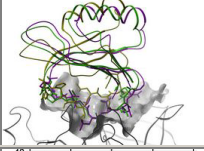
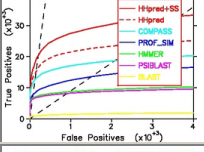
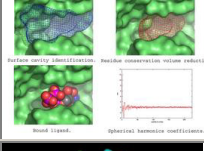
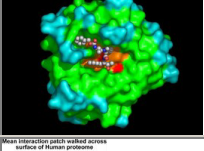
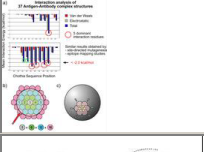
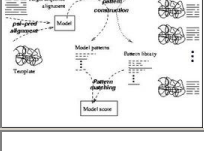
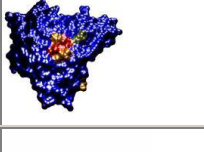
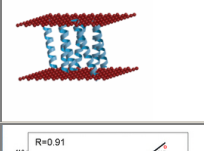
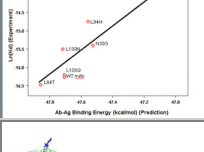

The Mean Field (MF) method is one of the most successful methods in CPD[6, 8-10]. It has the advantage over other methods of searching for an ensemble of protein conformations of global minimum *free energy*[8], as opposed to the single conformation of global minimum *energy* most other methods search for. However, it is generally not straightforward to incorporate an arbitrary solvation model into the MF method. We have previously succeeded in implementing an ASP solvation model into the MF method[6], but, although an improvement in the accuracy of the estimated free energies is observed, they are still not accurate enough for quantitative CPD. In the work we report here, we describe a means of correctly incorporating the considerably more accurate Langevin Dipoles (LD) solvation model[11] into the MF framework for CPD. The LD solvation model is one of the most accurate quantitative solvation models currently available, and one of the most commonly applied in biochemical problems. This work represents a major step forward along the path toward CPD methods based

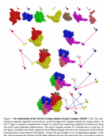
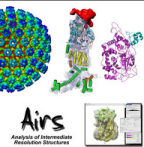
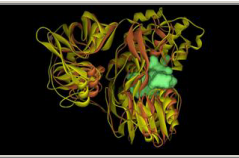
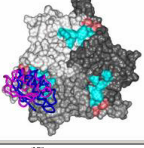
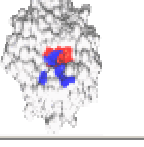

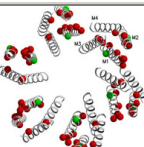
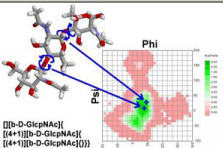


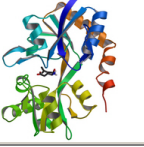
solely on first principles and should lead to a quantitative step forward in state-of-the-art CPD methodologies.

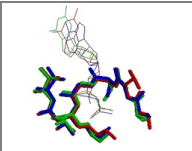
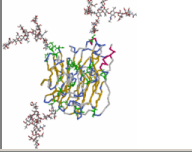
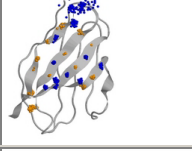
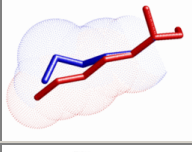
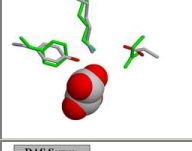
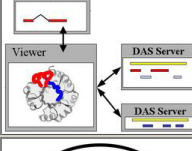
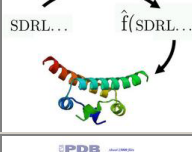
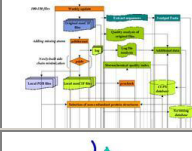

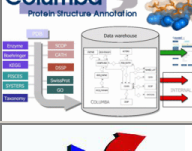
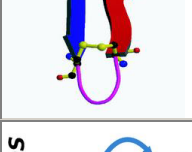
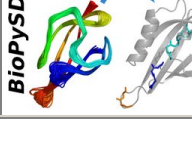
References:

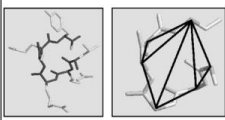
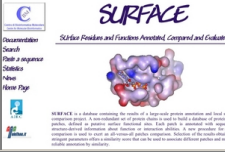
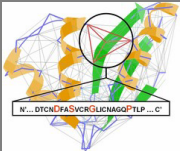
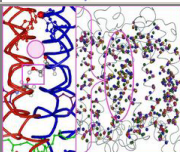

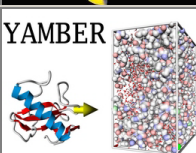
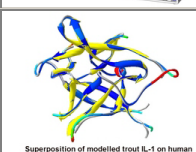
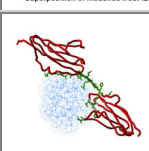

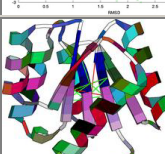
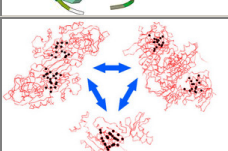
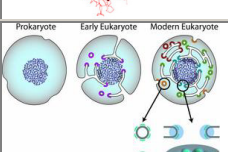
1. Davis, M.E. and J.A. McCammon, Electrostatics in Biomolecular Structure and Dynamics. *Chem. Rev.*, 1990. 90(3): p. 509-521.
 2. Smith, P.E. and B.M. Pettitt, Modeling Solvent in Biomolecular Systems. *J. Phys. Chem.*, 1994. 98(39): p. 9700-9711.
 3. Warshel, A. and A. Papazyan, Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.*, 1998. 8(2): p. 211-217.
 4. Cramer, C.J. and D.G. Truhlar, Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.*, 1999. 99(8): p. 2161-2200.
 5. Eisenberg, D. and A.D. McLachlan, Solvation energy in protein folding and binding. *Nature*, 1986. 319(6050): p. 199-203.
 6. Mendes, J., et al., Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants. *J. Comput.-Aided Mol. Des.*, 2001. 15(8): p. 721-740.
 7. Mendes, J., R. Guerois, and L. Serrano, Energy estimation in protein design. *Curr. Opin. Struct. Biol.*, 2002. 12(4): p. 441-446.
 8. Mendes, J., C.M. Soares, and M.A. Carrondo, Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction. *Biopolymers*, 1999. 50(2): p. 111-131.
 9. Mendes, J., et al., Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins*, 1999. 37(4): p. 530-543.
 10. de la Paz, M.L., et al., Computer-aided design of beta-sheet peptides. *J. Mol. Biol.*, 2001. 312(1): p. 229-246.
 11. Warshel, A. and S.T. Russell, Calculations of electrostatic interactions in biological systems and in solutions. *Q. Rev. Biophys.*, 1984. 17(3): p. 283-422.
-

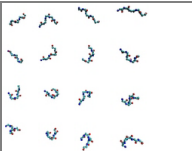
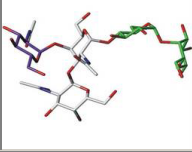
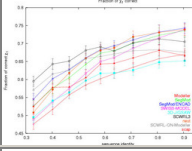
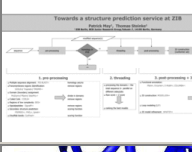
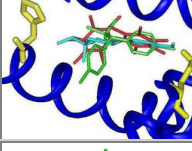
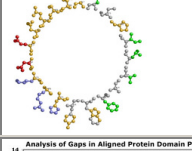
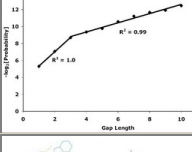
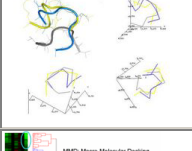
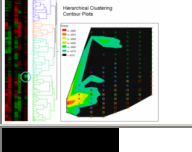
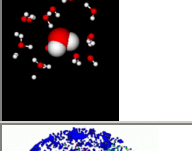
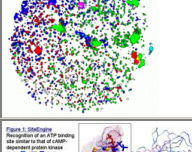
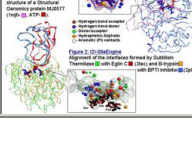
No.	Representative Figure	Authors	Title	Page
Invited Speaker Abstracts				
Invited-1		Robert B. Russell	A Structural Perspective on Protein Interactions & Complexes	9
Invited-2		Ron Unger	Using Accumulated Dot Matrices to Detect RNA and Protein Structures	9
Invited-3		Gregorio Fernandez, Christine Kiel, Luis Serrano	In silico Prediction of Protein-Protein Interactions: How and When It Can Work	10
Invited-4		Gunnar von Heijne	Genome-wide Membrane Protein Topology Predictions	10
Invited-5		Florencio Pazos, Michael JE Sternberg	Automated Prediction of Protein Function from Structure	10
Invited-7		Martin Madera, Julian Gough, Sarah Kummerfeld, Cyrus Chothia ³	The Homology of Genome Sequences: Profile-Profile Sequence Matching and the SUPERFAMILY Database	11
Invited-8		Song Yang, Russell Doolittle, Philip E. Bourne	The Study of Evolution through Protein Domain Structure	11
Speaker Abstracts				
5		Rune Linding, Robert B. Russell, Lars J. Jensen, Francesca Diella, Peer Bork, Toby J. Gibson	Intrinsic Protein Disorder - Function, Disease and Structural Proteomics.	12
9		Barry J. Grant, Leo S. Caves, Rob Cross	Stripping down the Kinesin Molecular Motor: a Combined Informatics and Simulation Approach	12
15		Eran Eyal, Marvin Edelman, Vladimir Sobolev	The Influence of Crystal Packing on Protein structure	13
18		Andreas Heger, Liisa Holm	Fast Fold Recognition by Consensus Alignment within Transitive Closure	14
22		Oxana V. Galzitskaya, Sergiy O. Garbuzynskiy, Bogdan S. Melnik, Michail Y. Lobanov, Alexei V. Finkelstein	Comparison of X-ray and NMR Protein Structures	15
24		Sanne Abeln, Charlotte M. Deane	Investigating Protein Structure Evolution by Fold Usage on Genomes	16

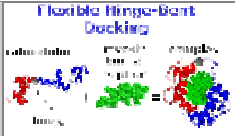
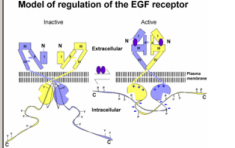
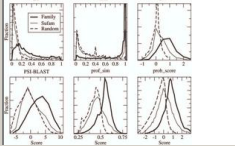


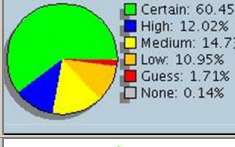
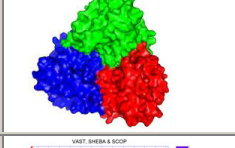
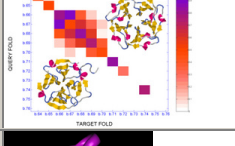
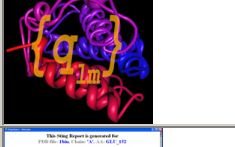

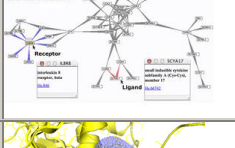
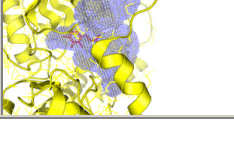
28		Antonio Del Sol, Paul A. O'Meara	Small-World Network Approach to Identify Key Residues in Protein-Protein Interaction.	17
31		Maxim Chapovalov, Roland L. Dunbrack	Using Statistical Analysis of Electron Density to Evaluate Protein Side-Chain Conformations and Rotamer Disorder	18
36		Maxim Totrov	Protein-Protein Docking Simulations with Local Backbone Flexibility	19
41		Johannes Soeding, Andrei N. Lupas	Homology Search By HMM-HMM Comparison Detects More Than Three Times As Many Remote Homologs as PSIBLAST or HMMER	21
43		Richard J. Morris, Abdullah Kahraman, Rafael Najmanovich, Fabian Glaser, Roman Laskowski, Janet M. Thornton	Protein Active Site Identification and Fast Shape Comparison	22
45		Arye Shemesh, Gil Amitai, Einat Sitbon, Maxim Shklar, Dvir Netaneli, Ilya Venger, Shmuel Pietrokovski	Structural Analysis of Residue Interaction Graphs	22
46		Gestel David, Ian Humphery-Smith	The Binding-Site Diversity of the Entire Non-denatured, Surface-exposed Human Proteome	23
47		Inge Jonassen, William R. Taylor	SPREK: A Method for Evaluating Structural Models using Structural Patterns	24
53		Jaspreet S. Sodhi, Kevin Bryson, Liam J. McGuffin, Jonathan J. Ward, Lornenz Wernisch, David T. Jones	Predicting the Location and Identity of Protein Functional Sites, Application to Genomic Fold Recognition	24
68		Marialisa Pellegrini-Calace, William R. Taylor, David T. Jones	FILM2: a Novel Method for the Prediction of Transmembrane Helical Bundles in a Membrane-like Environment	25
72		Qiaojuan Jane Su, Orit Foord, Scott Klakamp, Holly Tao, Larry L. Green	Modeling and Simulation of Antibody-Antigen Interaction: An Integrated Approach	26
74		Joerg Hackermueller, Nicole-Claudia Meisner, Manfred Auer, Markus Jaritz	mRNA Openers – Computationally Designed Modulators of mRNA Secondary Structure which Manipulate Gene Expression	27


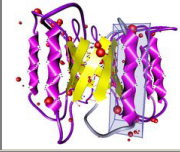
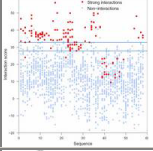

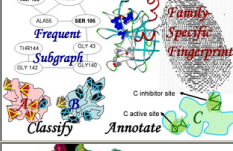
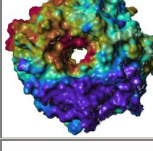
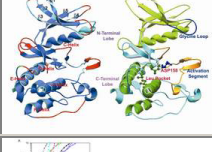
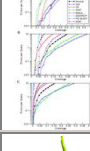
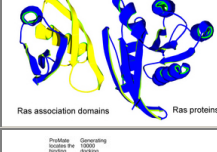
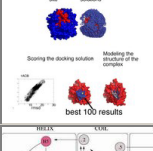
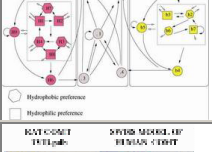
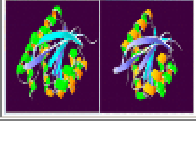
76		Yuval Inbar, Haim J. Wolfson, Ruth Nussinov	Prediction of Multi-Molecular Assemblies by Multiple Combinatorial Docking.	28
77		Matthew L. Baker, Wah Chiu	Analysis of Intermediate Resolution Structures	29
Laptop Session Abstracts				
1		Anne Marie Quinn, Luke Fisher, Dana Haley-Vicente	From Gene to Function: In Silico Warfare on the West Nile Virus	30
2		Efrat Ben-Zeev, Miriam Eisenstein	Weighted Geometric Docking with MolFit: Incorporating External Information in the Rotation-Translation Scan	31
3		James Bradford, David Westhead	Prediction of Protein-Protein Binding Sites using Support Vector Machines	32
4		Kim Henrick, Tom Oldfield, Adel Golovin, John Tate, Sameer Velankar, Harry Boutselakis, Dimitris Dimitropoulos, Peter Keller, Eugene Krissinel, Phil McNeil, Jorge Pineda, Abdelkrim Rachedi, Antonio Suarez-Uruena, Jawahar Swaminathan, Mohamed Tagari	MSD Relational Database, Search and Visualisation of Queries and Results	32
6		Sarel J. Fleishman, Vinzenz M. Unger, Mark Yeager, Nir Ben-Tal	A Model Structure of the Gap-Junction Transmembrane Domain Specifying C-alpha Positions	33
7		Thomas Lütke, Claus W. von der Lieth	Towards Deciphering the Structural Features of Glycosylation Sites: Analysis of Carbohydrate-Related Data contained in the Protein Data Bank	33
8		Wolfgang F. Bluhm on behalf of the PDB team	PDB-in-a-Box	35
10		Howard J. Feldman, Christopher WV Hogue, Kevin Snyder	MMDBBIND - Macromolecular Interactions with Atomic Level Detail	36
11		Philip C. Biggin, Yalini Arinaminpathy, Mark SP Sansom	Conformational Dynamics of Glutamate Receptors: Simulation Studies.	37

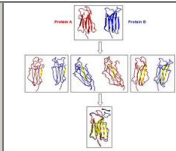
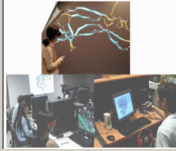
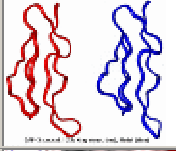
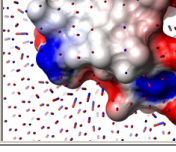
12		Nicola D. Gold, Richard M. Jackson	A Searchable Database for Comparing Protein-Ligand Binding Sites for the Discovery of Structure-Function Relationships	37
13		Andreas Bohne-Lang, Claus W. von der Lieth	Glyprot: a Web-tool for in-silico Glycosylation of Proteins	38
14		Vladimir Potapov, Vladimir Sobolev, Marvin Edelman, Alexander Kister, Israel Gelfand	Interface Cores in Sandwich-like Proteins	40
16		Eran Eyal, Vladimir Sobolev, Rafael Najmanovich, Brendan J. McConkey, Marvin Edelman	Modeling Side Chain Conformations using Contact Surfaces and solvent Accessible Surface	40
17		Gail J. Bartlett, James W. Torrance, Craig T. Porter, Jonathan A. Barker, Alex Gutteridge, Malcolm W. MacArthur, Janet M. Thornton	Generation of 3D Templates for Enzymes in the Catalytic Site Atlas: Analysis of Catalytic Residue Geometry and Utility of Automatically-generated templates for Function Prediction from 3D Structure	42
19		Andreas Prlic, Thomas A. Down, Tim JP. Hubbard	Protein DAS, Distributed Annotation System for Proteins	43
20		Ingolf E. Sommer, Joerg Rahnenfuehrer, Francisco S. Domingues, Ulrik de Lichtenberg, Thomas Lengauer	Predicting Protein Structure Classes from Function Predictions	43
21		Jean-Luc Pellequer, Shu-wen W. Chen, Gilles Imbert, Olivier Pible, Isabelle Vergely	Selecting Non-Redundant Protein Structures from the Protein Data Bank	44
23		Samantha L. Kaye, Mark SP Sansom, Philip C. Biggin	Simulation and Modelling studies of the NMDA Receptor	45
25		Silke Trissl, Kristian Rother, Ulf Leser	COLUMBA - A Database of Annotated Protein Structures	46
26		Merridee A. Wouters, Ken K. Lau, Philip J. Hogg	Cross-Strand Disulphides in Cell Entry Proteins: Redox Switches in Protein Nanomachines	47
27		Alessandro Pandini, Giancarlo Mauri, Laura Bonati	BioPySDS: an Object-Oriented Interface to Manage Protein Dynamics in an Evolutionary Framework	49

29		Chen Yanover, Ori Shachar, Yair Weiss	Inference in Graphical Models for Side-chain Prediction	50
30		Fabrizio Ferre', Gabriele Ausiello, Manuela Helmer-Citterich, Andreas Zanzoni	Large Scale Surface Comparison for the Identification of Functional Similarities in Unrelated Proteins	51
32		Ruchir R. Shah, Luke Huan, Deepak Bandyopadhyay, Wei Wang, Alexander Tropsha	Structure Based Identification of Protein Family Signatures for Function Annotation	51
33		Ilan Samish, Eran Goldberg, Oksana Kerner, Avigdor Scherz	Centrality of Weak Interhelical H-bonds in Membrane Protein Functional Assembly and Conformational Gating	53
34		Dmitry S. Kanibolotsky, Konstantin A. Odynets, Alexander I. Kornelyuk	Homology Modeling and Molecular Dynamics Simulation Study of Cytokine-like C-terminal Module of Mammalian Tyrosyl-tRNA Synthetase	53
35		Elmar Krieger, Tom Darden, Sander B. Nabuurs, Alexei V. Finkelstein, Gert Vriend	Making Optimal Use of Empirical energy Functions: Force Field Parameterization in crystal Space	54
37		Antonis Koussounadis, David W Ritchie, Antonis Koussounadis, Chris J. Secombes	Modelling b-Trefoil Proteins Using an Object-Oriented Database	55
38		Juan Fernandez-Recio, Tom L. Blundell	Computer Simulations of Protein-Protein Interactions: Rigid-Body Docking and Flexible-Link Refinement Applied to Multi-Domain Signalling Complexes	55
39		Hugh P. Shanahan, Sue Jones, Carles Ferrer, Janet M. Thornton	Detecting DNA-binding Proteins using Structural Motifs and Electrostatics	56
40		Robert M. MacCallum	Self-Organized Maps of Sequence Profiles for Protein Structure Prediction	57
42		Craig J. Lucas, Andrew Bulpitt	Automatic Identification of Key Patterns Within Multiple Protein Structures	58
44		Damien Devos, Svetlana Dokudovskaya, Marc A. Marti-Renom, Rosemary Williams, Brian T. Chait, Andrej Sali, Michael P. Rout	Evidence for A Common Evolutionary Origin of Nuclear Pore Complexes and Coated Vesicles	60

48		Cristina Benros, Alexandre G. de Brevern, Serge Hazout	Predicting Local Structural Candidates from Sequence by the "Hybrid Protein Model" Approach	61
49		Abraham Nahmany, Francesco Strino, Jimmy Rosen, Graham J.L. Kemp, Per-Georg Nyholm	Prediction of Oligosaccharide Conformations using Genetic Algorithms and Molecular Mechanics MM3	62
50		Björn Wallner, Arne Elofsson	Benchmark of Different Homology Modelling Programs	63
51		Patrick May, Thomas Steinke, Michael Meyer	THESEUS: A Parallel Threading Core	63
52		Ermanna Rovida, Pasqualina D'Ursi, Paola Fossa, Luciano Milanesi	Modelling the Interaction of Human Estrogen Receptor alpha with Organic Polychlorinated Compounds	64
54		Julian Mintseris, Zhiping Weng	Optimized Protein Representations from Information Theory	65
55		Nalin C. Goonesekere, Byungkook Lee	Frequency of Gaps Observed in a Structurally Aligned Protein Pair Database Suggests a SIMPLE GAP PENALTY FUNCTION	65
56		Victoria F. Dominguez Del Angel, Luis Mochan, Bernard Caudron, Charles Roth	PP2A(alpha) Interactions Domains, a Model of 3D Pattern Representation	66
57		Qingjuan Gu, William W. Reenstra, John J. Rux	MMD: A Macro-Molecular Docking Program for Locating Domain-Domain Interactions	67
58		Gianni De Fabritiis, Rafael Delgado-Buscalioni, P. V. Coveney	A Minimisation Method for the efficient Insertion of Solvent Water Molecules	67
59		Jim Procter, Andrew Torda	Protein sequence-structure Alignments for Prediction, Classification, and Visualization	68
60		Alexandra Shulman-Peleg, Shira Mintz, Ruth Nussinov, Haim J. Wolfson	Protein Functional Sites Recognition and Classification	69

61		Dina Schneidman, Yuval Inbar, Ruth Nussinov, Haim J. Wolfson	FlexDock: An Algorithm for Flexible Hinge-Bent Docking.	71
62		Meytal Landau, Sarel J. Fleishman, Nir Ben-Tal	Down-Regulation of the Catalytic Activity of the EGF Receptor via Direct Contact between the Kinase and C-Terminal Domains	72
63		Håkan Viklund, Arne Elofsson	Profile-profile HMMs for Structure Predictions	73
64		Thomas P. Walsh, Geoffrey J. Barton	A PERL Library for Generating Hierarchical, Tree-based Browser Interfaces for SCOP Database Searches.	74
65		Jan Reichert, Kristina Mehliß, Juergen Suehnel	The IMB Jena Image Library of Biological Macromolecules - Recent Developments	75
66		Liam J. McGuffin, Stefano A. Street, David T. Jones, Søren-Aksel Sørensen	The Genomic Threading Database	76
67		Hannes Ponstingl, Janet M. Thornton	Structure and Sequence Characteristics of Oligomeric Proteins	77
69		Vichetra Sam, Chin-Hsien (Emily) Tai, Peter J. Munson, Jean Garnier, Jean-François Gibrat, Byungkook Lee	Comparison of Protein Structural Comparison Methods, VAST and SHEBA, with the SCOP Classification, using Statistical Methods	77
70		Apostol Gramada, Philip E. Bourne	A Multipole-Based Method for Comparison of Protein Structures	78
71		Goran Neshich, Roberto Higa, Adaauto Mancini, Paula Kuzer, Michel Yamagishi, Renato Fileto	Selecting a Set of the Protein Structure Descriptors Uniquely Defining the Active Site Residues	79
73		Javier De Las Rivas, Carlos Prieto, Alberto De Luis	Exploring Protein Interaction Networks and Functional Relationships at a Genomic Scale with a Agile Browser	80
75		Rafael Najmanovich, Richard J. Morris, Roman Laskowski, Janet M. Thornton	Shape Description of Protein Clefts	81

78		Murad Nayaal, Barry Honig	Protein Surface Cavities and their Role in Protein-Protein Interactions	81
79		Silvia N. Crivelli, James Lu, James Lu, Oliver Kreylos, Nelson Max, and Wes Bethel	ProteinShop: a Tool for Interactive Protein Manipulation	82
80		Jessica Fong, Amy Keating, Mona Singh	Large Scale, High-Confidence Predictions of bZIP Protein-Protein Interactions	83
81		Jessica C. Shapiro, Douglas L. Brutlag	Automatic Structural Motif Discovery using FOLDMINER	83
82		Deepak Bandyopadhyay, Luke Huan, Wei Wang, Jack Snoeyink, Jan Prins, Alexander Tropsha	Graph Representations and Algorithms for Protein Family Classification and Functional Annotation	84
83		Agnes Tan	Automated Construction of WD40 proteins	86
84		Alvin Ng	Homology Model and Substrate Recognition of Yeast Prk1p Kinase	86
85		Michael L. Sierk, William R. Pearson	Benchmarking Protein Structure Alignment Statistics Against the CATH Database	88
86		Christina Kiel, Luis Serrano, Alfred Wittinghofer, Sabine Wohlgemuth	Recognizing and Defining true Ras Association (RA) Domains based on Protein Sequence and Structure	88
87		Kay Gottschalk, Hani Neuvirth, Gideon Schreiber	Docking Protein Complexes from their Predicted Binding Site	89
88		Juliette Martin, Jean-François Gibrat, François Rodolphe	Hidden Markov Model for Protein Secondary Structure	90
89		C. Jayaprabha, N. Santhi, J. Kamesh, P. Chitra, N. Bharathi, S. Krishnaveni, S. Valarmathi, S. Sujatha and S. Seethalakshmi	<i>In Silico</i> Transfer of Neurotransmitter Transporter Motif Between Structurally Analogous Protein (Catechol-O-methyltransferase)	91

90		Giacomo De Mori, Raul Mendez, Didier Croes	Analysis of Conformational Changes on Protein - Protein Complexes.	92
91		YY Cai, BF Lu, ZW Fan, and KT Lim	3D Structure Image Generator using Virtual Reality Technology for Protein Immersive Visualization	92
92		Dinesh C. Soares, Paul N. Barlow, Dietlind L. Gerloff	The CCP-Module Model Database - Automated Large-Scale Protein Structure Modelling of Individual Human Complement Control Protein (CCP)-Modules	93
93		Joaquim Mendes, Luis Serrano	Incorporating an Accurate Solvation Model into Computational Protein Design	94

List of Authors

Format: **Family name, First name** Page No. (Abstract No.)

Abeln, Sanne 16(24)	Eisenstein, Miriam 31(2)	Keating, Amy 83(80)
Amitai, Gil 22(45)	Elofsson, Arne 63(50), 73(63)	Keller, Peter 32(4)
And, S. Sujatha 91(89)	Eyal, Eran 13(15), 40(16)	Kemp, Graham J.L. 62(49)
Arinaminpathy, Yalini 37(11)	Fan, ZW 92(91)	Kerner, Oksana 53(33)
Auer, Manfred 27(74)	Feldman, Howard J. 36(10)	Kiel, Christina 88(86)
Ausiello, Gabriele 51(30)	Fernandez, Gregorio 10(Invited-3)	Kiel, Christine 10(Invited-3)
Baker, Matthew L. 29(77)	Fernandez-Recio, Juan 55(38)	Kister, Alexander 40(14)
Bandyopadhyay, Deepak 84(82), 51(32)	Ferre', Fabrizio 51(30)	Klakamp, Scott 26(72)
Barker, Jonathan A. 42(17)	Ferrer, Carles 56(39)	Kornelyuk, Alexander I. 53(34)
Barlow, Paul N. 93(92)	Fileto, Renato 79(71)	Koussounadis, Antonis 55(37), 55(37)
Bartlett, Gail J. 42(17)	Finkelstein, Alexei V. 15(22), 54(35)	Kreylos, Oliver 82(79)
Barton, Geoffrey J. 74(64)	Fisher, Luke 30(1)	Krieger, Elmar 54(35)
Ben-Tal, Nir 72(62), 33(6)	Fleishman, Sarel J. 33(6), 72(62)	Krishnaveni, S. 91(89)
Ben-Zeev, Efrat 31(2)	Fong, Jessica 83(80)	Krissinel, Eugene 32(4)
Benros, Cristina 61(48)	Foord, Orit 26(72)	Kummerfeld, Sarah 11(Invited-7)
Bethel, And Wes 82(79)	Fossa, Paola 64(52)	Kuzer, Paula 79(71)
Bharathi, N. 91(89)	Galzitskaya, Oxana V. 15(22)	Landau, Meytal 72(62)
Biggin, Philip C. 37(11), 45(23)	Garbuzynskiy, Sergiy O. 15(22)	Laskowski, Roman 22(43), 81(75)
Bluhm, Wolfgang F. 35(8)	Garnier, Jean 77(69)	Lau, Ken K. 47(26)
Blundell, Tom L. 55(38)	Gelfand, Israel 40(14)	Lee, Byungkook 65(55), 77(69)
Bohne-Lang, Andreas 38(13)	Gerloff, Dietlind L. 93(92)	Lengauer, Thomas 43(20)
Bonati, Laura 49(27)	Gibrat, Jean-Francois 90(88), 77(69)	Leser, Ulf 46(25)
Bork, Peer 12(5)	Gibson, Toby J. 12(5)	Lichtenberg, Ulrik De 43(20)
Bourne, Philip E. 78(70), 11(Invited-8)	Glaser, Fabian 22(43)	Lim, And KT 92(91)
Boutselakis, Harry 32(4)	Gold, Nicola D. 37(12)	Linding, Rune 12(5)
Bradford, James 32(3)	Goldberg, Eran 53(33)	Lobanov, Michail Y. 15(22)
Brutlag, Douglas L. 83(81)	Golovin, Adel 32(4)	Lu, BF 92(91)
Bryson, Kevin 24(53)	Goonsekere, Nalin C. 65(55)	Lu, James 82(79), 82(79)
Bulpitt, Andrew 58(42)	Gottschalk, Kay 89(87)	Lucas, Craig J. 58(42)
Cai, YY 92(91)	Gough, Julian 11(Invited-7)	Luis, Alberto De 80(73)
Caudron, Bernard 66(56)	Gramada, Apostol 78(70)	Lupas, Andrei N. 21(41)
Caves, Leo S. 12(9)	Grant, Barry J. 12(9)	Lutke, Thomas 34(7)
Chait, Brian T. 60(44)	Green, Larry L. 26(72)	MacArthur, Malcolm W. 42(17)
Chapovalov, Maxim 18(31)	Gu, Qingjuan 67(57)	MacCallum, Robert M. 57(40)
Chen, Shu-wen W. 44(21)	Gutteridge, Alex 42(17)	Madera, Martin 11(Invited-7)
Chitra, P. 91(89)	Hackermueller, Joerg 27(74)	Mancini, Adauto 79(71)
Chiu, Wah 29(77)	Haley-Vicente, Dana 30(1)	Marti-Renom, Marc A. 60(44)
Chothia, Cyrus 11(Invited-7)	Hazout, Serge 61(48)	Martin, Juliette 90(88)
Coveney, P. V. 67(58)	Heger, Andreas 14(18)	Mauri, Giancarlo 49(27)
Crivelli, Silvia N. 82(79)	Helmer-Citterich, Manuela 51(30)	Max, Nelson 82(79)
Croes, Didier 92(90)	Henrick, Kim 32(4)	May, Patrick 63(51)
Cross, Rob 12(9)	Higa, Roberto 79(71)	McConkey, Brendan J. 40(16)
D'Ursi, Pasqualina 64(52)	Hogg, Philip J. 47(26)	McGuffin, Liam J. 76(66), 24(53)
Darden, Tom 54(35)	Hogue, Christopher WV 36(10)	McNeil, Phil 32(4)
David, Gestel 23(46)	Holm, Liisa 14(18)	Mehliss, Kristina 75(65)
De Brevern, Alexandre G. 61(48)	Honig, Barry 81(78)	Meisner, Nicole-Claudia 27(74)
De Fabritiis, Gianni 67(58)	Huan, Luke 51(32), 84(82)	Melnik, Bogdan S. 15(22)
De Las Rivas, Javier 80(73)	Hubbard, Tim JP. 43(19)	Mendes, Joaquim 94(93)
De Mori, Giacomo 92(90)	Humphery-Smith, Ian 23(46)	Mendez, Raul 92(90)
Deane, Charlotte M. 16(24)	Imbert, Gilles 44(21)	Meyer, Michael 63(51)
Del Sol, Antonio 17(28)	Inbar, Yuval 71(61), 28(76)	Milanesi, Luciano 64(52)
Delgado-Buscalioni, Rafael 67(58)	Jackson, Richard M. 37(12)	Mintseris, Julian 65(54)
Devos, Damien 60(44)	Jaritz, Markus 27(74)	Mintz, Shira 69(60)
Diella, Francesca 12(5)	Jayaprabha, C. 91(89)	Mochan, Luis 66(56)
Dimitropoulos, Dimitris 32(4)	Jensen, Lars J. 12(5)	Morris, Richard J. 22(43), 81(75)
Dokudovskaya, Svetlana 60(44)	Jonassen, Inge 24(47)	Munson, Peter J. 77(69)
Domingues, Francisco S. 43(20)	Jones, David T. 76(66)	Nabuurs, Sander B. 54(35)
Dominguez Del Angel, Victoria F. 66(56)	Jones, Sue 56(39)	Nahmany, Abraham 62(49)
Doolittle, Russell 11(Invited-8)	Jones, David T. 24(53), 25(68)	Najmanovich, Rafael 81(75), 22(43), 40(16)
Down, Thomas A. 43(19)	Kahraman, Abdullah 22(43)	Nayal, Murad 81(78)
Dunbrack, Roland L. 18(31)	Kamesh, J. 91(89)	Neshich, Goran 79(71)
Edelman, Marvin 13(15), 40(14), 40(16)	Kanibolotsky, Dmitry S. 53(34)	Netanel, Dvir 22(45)
	Kaye, Samantha L. 45(23)	Neuvirth, Hani 89(87)

Ng, Alvin 86(84)
 Nussinov, Ruth 69(60), 71(61), 28(76)
 Nyholm, Per-Georg 62(49)
 O'Meara, Paul A. 17(28)
 Odynets, Konstantin A. 53(34)
 Oldfield, Tom 32(4)
 Pandini, Alessandro 49(27)
 Pazos, Florencio 10(Invited-5)
 Pearson, William R. 88(85)
 Pellegrini-Calace, Marialuisa 25(68)
 Pellequer, Jean-Luc 44(21)
 Pible, Olivier 44(21)
 Pietrokovski, Shmuel 22(45)
 Pineda, Jorge 32(4)
 Ponstingl, Hannes 77(67)
 Porter, Craig T. 42(17)
 Potapov, Vladimir 40(14)
 Prieto, Carlos 80(73)
 Prins, Jan 84(82)
 Prlic, Andreas 43(19)
 Procter, Jim 68(59)
 Quinn, Anne Marie 30(1)
 Rachedi, Abdelkrim 32(4)
 Rahnenfuehrer, Joerg 43(20)
 Reenstra, William W. 67(57)
 Reichert, Jan 75(65)
 Ritchie, David W 55(37)
 Rodolphe, Frannois 90(88)
 Rosen, Jimmy 62(49)
 Roth, Charles 66(56)
 Rother, Kristian 46(25)
 Rout, Michael P. 60(44)
 Rovida, Ermanna 64(52)
 Russell, Robert B. 12(5), 9(Invited-1)
 Rux, John J. 67(57)
 Sali, Andrej 60(44)
 Sam, Vichetra 77(69)
 Samish, Ilan 53(33)
 Sansom, Mark SP 45(23), 37(11)
 Santhi, N. 91(89)
 Scherz, Avigdor 53(33)
 Schneidman, Dina 71(61)
 Schreiber, Gideon 89(87)
 Secombes, Chris J. 55(37)
 Serrano, Luis 88(86), 10(Invited-3), 94(93)
 Shachar, Ori 50(29)
 Shah, Ruchir R. 51(32)
 Shanahan, Hugh P. 56(39)
 Shapiro, Jessica C. 83(81)
 Shemesh, Arye 22(45)
 Shklar, Maxim 22(45)
 Shulman-Peleg, Alexandra 69(60)
 Sierk, Michael L. 88(85)
 Singh, Mona 83(80)
 Sitbon, Einat 22(45)
 Snoeyink, Jack 84(82)
 Snyder, Kevin 36(10)
 Soares, Dinesh C. 93(92)
 Sobolev, Vladimir 40(14), 40(16), 13(15)
 Sodhi, Jaspreet S. 24(53)
 Soeding, Johannes 21(41)
 Sommer, Ingolf E. 43(20)
 Sorensen, Soren-Aksel 76(66)
 Steinke, Thomas 63(51)
 Sternberg, Michael JE 10(Invited-5)
 Street, Stefano A. 76(66)
 Strino, Francesco 62(49)
 Su, Qiaojuan Jane 26(72)
 Suarez-Uruena, Antonio 32(4)
 Suehnel, Juergen 75(65)
 Swaminathan, Jawahar 32(4)
 Tagari, Mohamed 32(4)
 Tai, Chin-Hsien (Emily) 77(69)
 Tan, Agnes 86(83)
 Tao, Holly 26(72)
 Tate, John 32(4)
 Taylor, William R. 24(47), 25(68)
 Thornton, Janet M. 77(67), 42(17), 56(39), 22(43), 81(75)
 Torda, Andrew 68(59)
 Torrance, James W. 42(17)
 Totrov, Maxim 19(36)
 Trissl, Silke 46(25)
 Tropsha, Alexander 51(32), 84(82)
 Unger, Ron 9(Invited-2)
 Unger, Vinzenz M. 33(6)
 Valarmathi, S. 91(89)
 Velankar, Sameer 32(4)
 Venger, Ilya 22(45)
 Vergely, Isabelle 44(21)
 Viklund, Hikan 73(63)
 Von Heijne, Gunnar 10(Invited-4)
 Von der Lieth, Claus W. 34(7), 38(13)
 Vriend, Gert 54(35)
 Wallner, Björn 63(50)
 Walsh, Thomas P. 74(64)
 Wang, Wei 51(32), 84(82)
 Ward, Jonathan J. 24(53)
 Weiss, Yair 50(29)
 Weng, Zhiping 65(54)
 Wernisch, Lornenz 24(53)
 Westhead, David 32(3)
 Williams, Rosemary 60(44)
 Wittinghofer, Alfred 88(86)
 Wohlgemuth, Sabine 88(86)
 Wolfson, Haim J. 28(76), 69(60), 71(61)
 Wouters, Merridee A. 47(26)
 Yamagishi, Michel 79(71)
 Yang, Song 11(Invited-8)
 Yanover, Chen 50(29)
 Yeager, Mark 33(6)
 Zanzoni, Andreas 51(30)

List of Participants

First	Family	Organization	Country	Email
Sanne	Abeln	Dept. of Statistics, U. of Oxford	UK	abeln@stats.ox.ac.uk
Vikram	Alva Kullanja	MPI for Developmental Bio	GERMANY	vikram.alva@tuebingen.mpg.de
Antonina	Andreeva	MRC CPE Cambridge	UK	tony@mrc-lmb.cam.ac.uk
Collins	Appiah	E.C.G.	GHANA	achezy@yahoo.com
Anna	Badimo	Wits U.	SOUTH AFRICA	anna@cs.wits.ac.za
Matthew	Baker	Baylor College of Medicine	USA	mbaker@bcm.tmc.edu
Deepak	Bandyopadhyay	Dept. of Computer Science U. North Carolina	USA	debug@cs.unc.edu
Gail	Bartlett	Imperial College London	UK	g.bartlett@imperial.ac.uk
Geoffrey	Barton	School of Life Sciences U. of Dundee	UK	geoff@compbio.dundee.ac.uk
Nicolas	Baurin	Aventis Pharma	FRANCE	nicolas.baurin@aventis.Com
Claude	Beazley	Birkbeck College U. of London	UK	cbleazley@rfcgr.mrc.ac.uk
Riccardo	Bennett-Lovsey	Imperial College, London	UK	riccardo.bennett-lovsey@imperial.ac.uk
Cristina	Benros	EBGM, INSERM E0346 U. Paris 7	FRANCE	benros@ebgm.jussieu.fr
Nir	BenTal	Tel Aviv U. Dept. of Biochemistry	ISRAEL	bental@ashtoret.tau.ac.il
Efart	Ben-Zeev	Weizmann Inst. of Science	ISRAEL	efrat.ben-zeev@weizmann.ac.il
Phil	Biggin	Oxford U.	UK	phil@biop.ox.ac.uk
Alan	Bleasby	Medical Research Council Rosalind Franklin Centre	UK	ableasby@rfcgr.mrc.ac.uk
Wolfgang	Bluhm	Protein Data Bank Univ. of Calif., San Diego	USA	mail@wbluhm.com
Andreas	Bohne-Lang	Central Spectroscopy German Cancer Research Center	GERMANY	a.bohne@dkfz.de
Steven	Bottomley	U. Curtin WABRI	AUSTRALIA	S.Bottomley@curtin.edu.au
Philip	Bourne	U. of California, San SDSC	USA	bourne@sdsc.edu
James	Bradford	U. Of Leeds	UK	bmbjrb@bmb.leeds.ac.uk
Douglas	Brutlag	Stanford U. Dept. of Biochemistry	USA	brutlag@stanford.edu
Nicholas	Burgoyne	U. of Leeds	UK	bmb9njb@bmb.leeds.ac.uk
Michael	Carson	U. of Alabama, Center for Biophysical Science	USA	carson@uab.edu
Leo	Caves	U. of York Dept. of Biology	UK	lsdc1@york.ac.uk
Meiping	Chang	Pfizer Inc.	USA	meiping.chang@pfizer.com
Shu-wen	Chen	ABC	FRANCE	cmft551@yahoo.com
Seung Joo	Cho	KIST	KOREA	chosj@kist.re.kr
Inhee	Choi	Ewha Womans U.	SOUTH KOREA	inheec@ewha.ac.kr
Cyrus	Chothia	MRC Laboratory of Molecular Biology	UK	chc1@mrc-lmb.cam.ac.uk
Manuel	Corpas	U. of Manchester	UK	mcorpas@cs.man.ac.uk
Silvia	Crivelli	Lawrence Berkeley National Lab	USA	SNCrivelli@lbl.gov
Tjaart	de Beer	Dept of Biochemistry U. of Pretoria	SOUTH AFRICA	tjaart@tuks.co.za
Gianni	De Fabritiis	Universilty College of London	UK	g.defabritiis@ucl.ac.uk
Javier	De Las Rivas	Centro Investigacion Cancer CSIC - USAL	SPAIN	jrvivas@usal.es
Antonio	Del Sol	Frontier Research Division, Fu	JAPAN	ao-mesa@fujirebio.co.jp
Damien	Devos	UCSF	USA	damien@salilab.org
Francisco	Domingues	Max-Planck-Institut Informatik Saarbruecken	GERMANY	doming@mpi-sb.mpg.de
Victoria	Dominguez Del Angel	Institut Pasteur - BBMI	FRANCE	victoria@pasteur.fr

First	Family	Organization	Country	Email
William	Duddy	U. of Glasgow	UK	bduddy@chem.gla.ac.uk
Roland	Dunbrack	Fox Chase Cancer Center	USA	RL_Dunbrack@fccc.edu
Miriam	Eisenstein	Weizmann Ins. of Science	ISRAEL	miriam.eisenstein@weizmann.ac.il
Khalil	El Mazouari	Ablynx	BELGIUM	khalil.elmazouari@ablynx.com
Arne	Elofsson	Stockholm Bioinformatics Centre Stockholm U.	SWEDEN	arne@sbcsu.se
Eran	Eyal	Weizmann Inst. of Science	ISRAEL	eran.eyal@weizmann.ac.il
Marcin	Feder	IIMCB Warsaw	POLAND	marcin@genesilico.pl
Howard	Feldman	The Blueprint Initiative Mount Sinai Hospital	CANADA	feldman@mshri.on.ca
Juan	Fernandez-Recio	U. of Cambridge Dept. of Biochemistry	UK	juan@cryst.bioc.cam.ac.uk
Sarel	Fleishman	Tel-Aviv U.	ISRAEL	sarel@post.tau.ac.il
Keiran	Fleming	Imperial College London Keiran Fleming	UK	k.fleming@imperial.ac.uk
Tor	Fli	MatNat/U. of Tromsø	NORWAY	tor@math.uit.no
Menachem	Fromer	Computer Science, Hebrew U. of Jerusalem	ISRAEL	fromer@cs.huji.ac.il
Michal	Gajda	Warsaw U. of Technology	POLAND	misiek@genesilico.pl
Oxana	Galzitskaya	Inst. of Protein Research	RUSSIA	ogalzit@vega.protres.ru
Richard	Gamblin		UK	bms8rjg@bmb.leeds.ac.uk
David	Gilbert	Bioinformatics Research Centre U. of Glasgow	UK	drq@brc.dcs.gla.ac.uk
Fiona Gwendolyn	Ginn Nielsen	U. of Southern Denmark Odense	DENMARK	iscb@glyn.dk
Apostol	Gramada	U. of California San Diego	USA	agramada@sdsc.edu
Barry	Grant	U. of York	UK	grant@ysbl.york.ac.uk
Markus	Gruber	Max-Planck-Inst. Tuebingen	GERMANY	markus.gruber@tuebingen.mpg.de
Joerg	Hackermueller	Novartis Inst. for Biomedical Research Vienna	AUSTRIA	joerg@tbi.univie.ac.at
Dana	Haley-Vicente	Accelrys	USA	dhv@accelrys.com
Manuela	Helmer-Citterich	Dept. Biology U. of Rome Tor Vergata	ITALY	citterich@uniroma2.it
Kim	Henrick	EMBL-EBI	UK	copeland@ebi.ac.uk
Annette	Hoglund	Simulation of Biological Systems	GERMANY	hoeglund@informatik.uni-tuebingen.de
Liisa	Holm	Inst. of Biotechnology U. of Helsinki	FINLAND	liisa.holm@helsinki.fi
Janet	Horrocks	U. of Abertay	UK	mltj@tay.ac.uk
Ian	Humphrey-Smith	Dept. of Pharmaceutical Proteomics, U. Utrecht	NETHERLANDS	ianhs@hotmail.com
Michael	Hurley	Medical Research Council Rosalind Franklin Centre	UK	mhurley@rfcgr.mrc.ac.uk
Rene	Hussong	Center for Bioinformatics Saarland U. Saarbrück	GERMANY	rene@bioinf.uni-sb.de
Yuval	Inbar	Tel Aviv U. School of Computer Science	ISRAEL	inbaryuv@tau.ac.il
Jon	Ison	Medical Research Council Rosalind Franklin Centre	UK	jison@rfcgr.mrc.ac.uk
Richard	Jackson	U. of Leeds	UK	jackson@bmb.leeds.ac.uk
Emily	Jefferson	Dundee U.	SCOTLAND	emily@compbio.dundee.ac.uk
Dr. Steven F.	Jennings	U. Arkansas MidSouth Bioinformatics Center	USA	sfjennings@ualr.edu
Nilofer	Jiwani	Aventis Pharmaceuticals	USA	nilofer.jiwani@aventis.com
Inge	Jonassen	Dept of Informatics/CBU U. of Bergen	NORWAY	inge@ii.uib.no
David	Jones	U. College London Dept. of Computer Science	UK	dtj@cs.ucl.ac.uk
Fourie	Joubert	U. of Pretoria Bioinformatics	SOUTH AFRICA	fjoubert@postino.up.ac.za
Quentin	Kaas	Institut de Genetique Humaine	FRANCE	kaas@ligm.igh.cnrs.fr
Alla	Karnovsky	Pfizer Inc.	USA	alla.karnovsky@pfizer.com

First	Family	Organization	Country	Email
Manjunatha	Karpenahalli	Max-Planck-Inst. (Developmental Biology)	GERMANY	manju@tuebingen.mpg.de
Kevin	Karplus	Univ. of California Santa Cruz	USA	karplus@soe.ucsc.edu
Samantha	Kaye	U. of Oxford Laboratory of Molecular Biophy	UK	samantha_l_kaye@yahoo.co.uk
Graham	Kemp	Chalmers U. of Technology	SWEDEN	kemp@cs.chalmers.se
Christina	Kiel	EMBL Heidelberg, Germany	GERMANY	kiel@embl.de
Juergen	Kopp	Biozentrum Basel & Swiss Inst. of Bioinformatics	SWITZERLAND	juergen.kopp@unibas.ch
Antonis	Koussounadis	Dept. of Computing S U. of Aberdeen	UK	akoussou@csd.abdn.ac.uk
Elmar	Krieger	Radboud U. Nijmegen	NETHERLANDS	elmar.krieger@cmbi.kun.nl
Eugene	Krissinel	EMBL-EBI	UK	copeland@ebi.ac.uk
Taejoon	Kwon	Samsung Advanced Inst. of Technology	SOUTH KOREA	taejoon_kwon@samsung.com
Wan	Kyu Kim	Medical Research Council Rosalind Franklin Centre	UK	wkim@rfcgr.mrc.ac.uk
Jon K	Laerdahl	CMBN, National Hospital Oslo	NORWAY	jonkl@usit.uio.no
Meytal	Landau	Tel-Aviv U.	ISRAEL	meytal@post.tau.ac.il
Byungkook	Lee	NIH NCI/CCR/LMB	USA	bk@nih.gov
Jessica S.	Liberles	Computational Biology Unit U. of Bergen	NORWAY	jessical@ii.uib.no
Rune	Linding	EMBL	GERMANY	linding@embl.de
Stefano	Lise	U. College London Dept. of Biochemistry & M	UK	lise@biochem.ucl.ac.uk
Binbin	Liu	U. of Leeds	UK	BLiu324@hotmail.com
Baifang	Lu	Nanyang Technological Universi	SINGAPORE	mbflu@ntu.edu.sg
Craig	Lucas	School of Computing U. of Leeds	UK	craigl@comp.leeds.ac.uk
Jimli	Luo	MRC Toxicology Unit	UK	jl19@le.ac.uk
Thomas	Lutteke	DKFZ Heidelberg	GERMANY	t.lutetteke@dkfz.de
Robert	MacCallum	Stockholm Bioinformatics Ctr. Stockholm U.	SWEDEN	maccallr@sbc.su.se
Martin	Madera	MRC Lab. of Mol. Biol.	UK	mm238@mrc-lmb.cam.ac.uk
Dennis	Madsen	Novo Nordisk	DENMARK	dnnm@novonordisk.com
Sally	Mardikian	U. of Leeds	UK	bmsam@bmb.leeds.ac.uk
Rachid	Maroun	Institut Pasteur	FRANCE	rmaroun@pasteur.fr
Juliette	Martin	Unit ¹ MIG INRA Jouy-en Josas	FRANCE	jumartin@jouy.inra.fr
Patrick	May	Zuse Inst. Berlin	GERMANY	patrick.may@zib.de
Raul	Mendez	SCMBB - ULB (BELGIQUE)	BELGIUM	raul@scmbb.ulb.ac.be
Julian	Mintseris	Boston U.	USA	julianm@bu.edu
Diego	Miranda-Saavedra	Bioinformatics Laboratory, Wel	UK	diego@compbio.dundee.ac.uk
Ernesto	Moreno	Center of Molecular Immunology	CUBA	emoreno@ict.cim.sld.cu
Hugh	Morgan	Birkbeck College U. of London	UK	hmorgan2@rfcgr.mrc.ac.uk
Richard	Morgan	New England Biolabs	USA	morgan@neb.com
Richard J.	Morris	EMBL-EBI European Bioinformatics Instit	UK	rjmorris@ebi.ac.uk
John	Moult	CARB, UMBI	USA	moult@umbi.umd.edu
Abraham	Nahmany	Gothenburg U. Medical Biochemistry	SWEDEN	avi_nahmany@yahoo.se
Rafael	Najmanovich	European Bioinformatics Inst. - EMBL EBI	UK	rafael.najmanovich@ebi.ac.uk
Stephane	Nauche	European Patent Office	NETHERLANDS	snauche@epo.org
Victor	Neduva	EMBL	GERMANY	neduva@embl.de
Goran	Neshich	EMBRAPA - CNPTIA	BRAZIL	neshich@cnptia.embrapa.br

First	Family	Organization	Country	Email
Alvin	Ng	Inst of Molecular & Cell Biolo	SINGAPORE	ngyj@imcb.a-star.edu.sg
John C.	Norvell	National Inst.s of Health	USA	norvellj@nigms.nih.gov
Maria	Novatchkova	I.M.P. Vienna	AUSTRIA	novatchkova@imp.univie.ac.at
Peter	Oledzki	U. of Leeds	UK	bmbpo@bmb.leeds.ac.uk
Alessandro	Pandini	Universit� degli Studi di Milano-Bicocca	ITALY	alessandro.pandini@unimib.it
Marialuisa	Pellegrini-Calace	EBI	UK	marial@ebi.ac.uk
Jean-Luc	Pellequer	CEA Valrho DSV/DIEP/SBTN	FRANCE	jp128444@atil.cea.fr
Shmuel	Pietrovovski	Weizmann Inst. of Science	ISRAEL	shmuel.pietrovovski@weizmann.ac.il
Jasmina	Ponjavic	Max-Planck-Inst.	GERMANY	jasmina.ponjavic@tuebingen.mpg.de
Hannes	Ponstingl	EMBL/EBI	UK	hpo@ebi.ac.uk
Vladimir	Potapov	Weizmann Inst. of Science	ISRAEL	vladimir.potapov@weizmann.ac.il
Andreas	Prlic	Wellcome Trust Sanger Inst.	UK	ap3@sanger.ac.uk
J B	Procter	ZBH U. of Hamburg	GERMANY	procter@zbh.uni-hamburg.de
Xueping	Quan	Edinburgh ICMB	UK	s0240137@sms.ed.ac.uk
Sathyapriya	Rajagopal	Indian Inst. of Science	INDIA	spriya@mbu.iisc.ernet.in
Gemma	Ravasio	DISAT - Universit� degli Studi di Milano-Bicocca	ITALY	gemmaravasio@tiscali.it
Ulf	Reimer	Jerini AG	GERMANY	reimer@jerini.com
Monika	Rella	U. of Leeds	UK	bmbmre@bmb.leeds.ac.uk
David	Ritchie	Computing Science Dept. U. of Aberdeen	UK	dritchie@csd.abdn.ac.uk
Walter	Rocchia	Scuola Normale Superiore NEST - PISA	ITALY	w.rocchia@sns.it
Ana	Rodrigues	U. of York	UK	rodrigues@ysbl.york.ac.uk
Alexander	Rodrigues	Inst. of Molecular Biology	UKRAINE	kornelyuk@imbg.org.ua
Torbjorn	Rognes	Centre for Mol Biol & Neurosci U. of Oslo	NORWAY	torbjorn.rognes@labmed.uio.no
Ermanna	Rovida	Inst. of Biomedical Technologies - CNR	ITALY	ermanna.rovida@itb.cnr.it
Rob	Russell	EMBL	GERMANY	russell@embl.de
John J.	Rux	Wistar Inst.	USA	rux@wistar.upenn.edu
Einar A.	R�dland	CMBN, Rikshospitalet Oslo	NORWAY	e.a.rodland@labmed.uio.no
Ilan	Samish	Weizmann Inst. of Science	ISRAEL	Ilan.Samish@weizmann.ac.il
Oliver	Sander	Max-Planck-Inst. for Informatics	GERMANY	osander@mpi-sb.mpg.de
Natchimuthu	Santhi	Sri Ramakrishna College of Art	INDIA	santhigowri@rediffmail.com
Alexander	Schleiffer	I.M.P. Vienna	AUSTRIA	schleiffer@imp.univie.ac.at
Ralf	Schmid	U. of Edinburgh School of Biology	UK	R.Schmid@ed.ac.uk
Dina	Schneidman	Tel Aviv U. School of Computer Science	ISRAEL	duhovka@tau.ac.il
Gideon	Schreiber	Weizmann Inst. of Science	ISRAEL	gideon.schreiber@weizmann.ac.il
Michael	Schroeder	Biotec TU Dresden	GERMANY	ms@mpi-cbg.de
Yakov	Sergeev	MPI for Developmental Biology	GERMANY	yakov.sergeev@tuebingen.mpg.de
Luis	Serrano	EMBL, Heidelberg, Germany	GERMANY	serrano@embl-heidelberg.de
Ruchir	Shah	U. of North Carolina at Chapel Hill	USA	ruchir@email.unc.edu
Hugh	Shanahan	EBI-EMBL Wellcome Trust Genome Campus	UK	Hugh.Shanahan@physics.org
Arye	Shemesh	Weizmann Inst. of Science	ISRAEL	arye.shemesh@weizmann.ac.il
Jennifer	Siepen	U. of Leeds	UK	bmbjrb@bmb.leeds.ac.uk
Michael	Sierk	U. of Virginia Biochemistry & Molecular Genet	USA	mls5w@virginia.edu

First	Family	Organization	Country	Email
Mona	Singh	Princeton U.	USA	mona@cs.princeton.edu
Vladimir	Sobolev	Weizmann Inst. of Science	ISRAEL	Vladimir.Sobolev@weizmann.ac.il
Jaspreet Singh	Sodhi	Bioinformatics Group U. College London	UK	j.sodhi@cs.ucl.ac.uk
Johannes	Soeding	MPI for Developmental Biology, Tuebingen	GERMANY	johannes.soeding@tuebingen.mpg.de
Tim-Robert	Soellick	Novartis Pharma AG	SWITZERLAND	tim-robert.soellick@pharma.novartis.com
Ingolf	Sommer	Max-Planck-Inst. for Informatics	GERMANY	sommer@mpi-sb.mpg.de
Alexander	Stark	EMBL Heidelberg	GERMANY	alex.stark@embl.de
Michael	Sternberg	Imperial College	UK	m.stenberg@ic.ac.uk
Francesco	Strino	Gothenburg U. Medical Biochemistry	SWEDEN	frs@interfree.it
Helena	Strombergsson	LCB Uppsala U.	SWEDEN	helenas@lcb.uu.se
Qiaojuan	Su	Abgenix	USA	jane.su@abgenix.com
Juergen	Suehnel	IMB Jena	GERMANY	jsuehnel@imb-jena.de
Agnes	Tan	Inst. of Molecular and Cel	SINGAPORE	mcbtanlc@imcb.a-star.edu.sg
Richard	Tan	Inst. of Molecular and Cel	SINGAPORE	rtan@imcb.a-star.edu.sg
Janet	Thornton	EMBL-EBI	UK	adams@ebi.ac.uk
Holmfridur	Thorsteinsdottir	Biozentrum Basel and Swiss Inst. of Bioinformatics	SWITZERLAND	holmfridur.thorsteinsdottir@unibas.ch
Steinar	Thorvaldsen	U. of Tromso	NORWAY	steinart@math.uit.no
Jean-Francois	Tomb	E.I.DuPont De Nemours & Co., I	USA	jean-francois.tomb@usa.dupont.com
James	Torrance	European Bioinformatics Inst.	UK	torrance@ebi.ac.uk
Maxim	Totrov	Molsoft	USA	max@molsoft.com
Silke	Trissl	Humboldt-Universitnt zu Berlin	GERMANY	trissl@informatik.hu-berlin.de
Katharina R.	Tufteland	Dep. of Molecular Biology U. of Bergen	NORWAY	katharina.tufteland@student.uib.no
Ron	Unger	Bar-Ilan U.	ISRAEL	ron@biocom1.ls.biu.ac.il
Alfonso	Valencia	National Centre for Biote CNB - CSIC	SPAIN	valencia@cnb.uam.es
Gunnar	von Heijne	Stockholm U.	SWEDEN	gunnar@dbb.su.se
Gert	Vriend	CMBI, U. of Nijmegen	NETHERLANDS	vriend@cmbi.kun.nl
Björn	Wallner	Stockholm Bioinformatics Cente Stockholm U.	SWEDEN	bjorn@sbc.su.se
Thomas	Walsh	School of Life Sciences U. of Dundee	UK	tom@compbio.dundee.ac.uk
Haim	Wolfson	School of Computer Science Tel Aviv U.	ISRAEL	wolfson@post.tau.ac.il
Monika	Wood	Promega Corporation	USA	mwood@promega.com
Merridee	Wouters	Victor Chang Cardiac Research Inst.	AUSTRALIA	m.wouters@victorchang.unsw.edu.au
Richard	Wroe	U. of Manchester	UK	finance@cs.man.ac.uk
Philip	Xiang	Roche Molecular Systems	USA	philip.xiang@roche.com
Weijia	Xu	U. of Texas, Dept. of Computer Science	USA	xwj@cs.utexas.edu
Lisa	Yan	Accelrys	USA	lly@accelrys.com
Song	Yang	Dept of Chem & Biochem UCSD	USA	soyang@ucsd.edu
Chen	Yanover	The Hebrew Univeristy	ISRAEL	cheny@cs.huji.ac.il
Piotr	Zielenkiewicz	Inst. Biochem. Biophys.	POLAND	piotr@ibb.waw.pl