



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

**Proceedings of  
3Dsig: The 2<sup>nd</sup> Structural Bioinformatics and  
Computational Biophysics Meeting –  
An ISMB Satellite Meeting**



**August 4-5, 2006**

**Fortaleza, Brazil**

SCIENTIFIC COMMITTEE: John Moult (Chair) - CARB, USA; Phil Bourne - UCSD, USA; Steven Brenner – Berkeley, USA, Joel Sussman – WIS, Israel, Janet Thornton - EBI, UK

ORGANIZING COMMITTEE: Ilan Samish – U. Penn, USA, Rafael Najmanovich – EBI, UK

CONSULTANT: Marvin Edelman – WIS, Israel

Sponsors:



<b>Table of Contents</b>	<b>Page</b>
Program	3
Abstracts	
Abstracts Table of Contents	6
Invited Speaker Abstracts	12
Speaker Abstracts	14
Laptop Session Abstracts	30
List of Participants	55

## Program for Day 1, August 4th 2006 Room A, Fortaleza Convention Center

9:00 *Opening remarks*

### Session 1: **Evolution** (Chair - [Ilan Samish](#))

9:05 [Nick V. Grishin](#). **Dramatic Evolutionary Changes in Protein Structures.**

9:35 Sandra Maguid, Gustavo Parisi, [Sebastian Fernandez-Alberti](#), Julian Echave. **Sequence-structure-flexibility relationship in protein evolution.**

9:55 Maria S. Fornasari, [Gustavo Parisi](#). **Protein Structure Diversity and Sequence Divergence in Evolution.**

10:15 [Diana K Ekman](#), Åsa K Björklund, Arne Elofsson. **Evolutionary History of Multi-Domain Proteins.**

10:35 *Coffee Break*

11:05 Christopher Dupont, Song Yang, Brian Palenik, [Philip E. Bourne](#). **Modern Proteomes Contain Imprints of Ancient Shifts in Environmental Chemistry.**

### Session 2: **Homology Modeling** (Chair - [Nick Grishin](#))

11:35 Francisco Melo, Alejandro Panjkovich, Andrej Sali, Marc A. Marti-Renom. **Structure Specific Statistical Potentials for Protein Structure Prediction.**

11:55 [Iris Antes](#), Thomas Lengauer. **DynaDock: Inhibitor based Refinement of Homology Modeled Protein Structures for Molecular Docking.**

12:15 *Lunch and Laptop Session 1*

15:00 **Discussion 1: The roles of Evolution, physics and informatics** (Chair - [Phil Bourne](#)).

### Session 3: **Ligand Binding** (Chair - [Steven Brenner](#))

15:45 [Rafael J. Najmanovich](#), Richard J. Morris, Janet M. Thornton. **Single-Family Analysis of Binding Site Structural Similarities.**

16:05 [Thomas Funkhouser](#), Roman Laskowski, Janet Thornton. **Matching Volumetric Models of Protein Active Sites.**

16:25 [Melissa R. Landon](#), Spencer C. Thiel, David R. Lancia Jr., Sandor Vajda. **Identification of Druggable "Hot Spots" in Ligand Binding Pockets by Computational Solvent Mapping of Proteins.**

16:45 *Coffee Break*

### Session 4: **Model Quality** (Chair - [John Moult](#))

17:15 Michal J. Gajda, Marta Kaczor, [Janusz M. Bujnicki](#), Anastasia Bakulina, Andrzej Kasparski, Alan M. Friedman, Chris Bailey-Kellogg. **FILTREST3D: a Simple Method for Discrimination of Multiple Structural Models Against Spatial Restraints.**

17:35 [Marcin Pawlowski](#), Janusz Bujnicki, Ryszard Matlak. **Meta-MQAP: an SVM-based meta-server for the quality assessment of protein models.**

17:55 [Wah Chiu](#), Matthew L Baker. **Modeling Protein Components of Biological Nano-Machines in Subnanometer Resolution CryoEM Density Maps.**

18:25 *Break: Busses to evening session*

20:00	<i>Get-together Dinner</i> at main headquarters hotel (Oasis Atlantic Imperial)
21:00	<u>John Moult</u> . <b>CASP: Progress, Bottlenecks, Prognosis, and Messages for Computational Biology</b>
23:00	<i>Lights out...</i>

## Program for Day 2, August 5th 2006

### Session 5: **Physics and Folding** (Chair - Goran Neshich), 9:30 - 10:30

9:00	<u>Jeffery G. Saven</u> . <b>Theoretical Methods for the Design and Engineering of Proteins.</b>
9:30	<u>Ilan Samish</u> , Oksana Kerner, David Kaftan, Eran Goldberg, Avigdor Scherz. <b>Datamining the Fourth Dimension from Crystal Structures: Structure-Function-Entropy Relationships in Membrane Proteins.</b>
9:50	<u>Anna C.V. Johansson</u> , Erik Lindahl. <b>Simulation of Amino Acid Solvation Structure in Transmembrane Helices and Implications for Membrane Protein Folding.</b>
10:10	<u>Ingo Ruczinski</u> , Kevin W. Plaxco, Tobin R. Sosnick. <b>On the Precision of Experimentally Determined Protein Folding Rates and [Phi]-Values.</b>
10:30	<i>Coffee Break</i>
11:00	<u>Janet M. Thornton</u> , Gemma L. Holliday, Alex Gutteridge, James Torrance, Gail J. Bartlett, Daniel E. Almonacid, Noel M.O'Boyle, Peter Murray-Rust, John B. O. Mitchell. <b>Using the Three Dimensional Structures of Enzymes to Understand Catalysis.</b> (invited presentation)

### Session 6: **Disorder** (Chair - Thomas Lengauer), 11:30 - 12:10

11:30	<u>Zsuzsanna Dosztányi</u> , Márk Sándor, István Simon. <b>Prediction of Order and Disorder at the Domain Level.</b>
11:50	<u>Avner Schlessinger</u> , Marco Punta, Burkhard Rost. <b>Identifying Natively Unstructured Proteins Using Residue Contact Predictions.</b>
12:10	<i>Lunch and Laptop Session 2</i>
15:00	<b>Discussion 2:</b> What are the greatest overlooked challenges in computational structural biology and what are the most over-studied low impact areas? (Chair: <u>Steven Brenner</u> )

### Session 7: **Protein-Protein Interactions and Domains** (Chair - Phil Bourne)

15:45	<u>Bingding Huang</u> , Michael Schroeder. <b>Integrate the Degree of Burial and Conservation of Surface Residues into Protein-protein Docking.</b>
16:05	<u>Stefano Lise</u> , David T Jones, Alice Walker-Taylor. <b>Docking protein domains in contact space.</b>
16:25	

16:45 *Coffee Break*

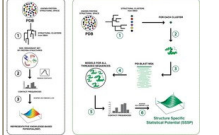
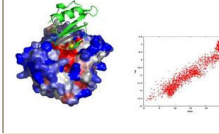
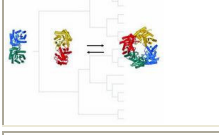
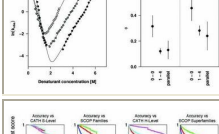
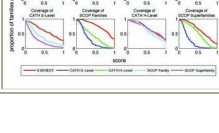
### Session 8: **Networks and Graphs** (Chair - Rafael Najmanovich)

17:15	<u>Frank P DiMaio</u> , Jude W Shavlik, George N. Phillips <b>A Probabilistic Approach to Protein Backbone Tracing in Electron Density Maps.</b>
-------	--

17:35	<u>Antonio del Sol</u> , Marcos J. Arauzo-Bravo, Dolors Amoros, Ruth Nussinov. <b>Network Robustness and Modularity of Protein Structures in the Identification of Key Residues for the Allosteric Communications.</b>
17:55	<u>Steven Brenner</u> . <b>The RNA Structural World.</b>
18:25	Closing remarks
18:30	<i>The End: Shuttle back to Hotels.</i>

## Abstracts Table of contents:


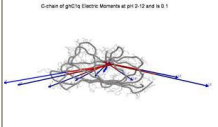

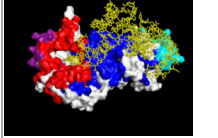
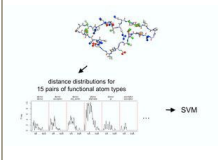
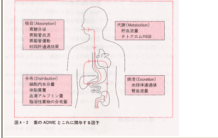
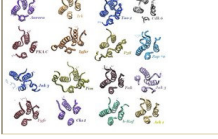
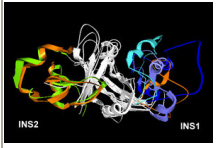

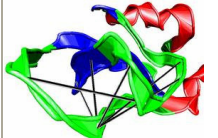
No	Representative Figure	Page	Authors	Title
<b>Invited Speaker Abstracts</b>				
<a href="#"><u>72</u></a>		Page 12	Nick V Grishin	Dramatic Evolutionary Changes in Protein Structures
<a href="#"><u>77</u></a>		Page 12	Christopher Dupont, Song Yang, Brian Palenik, Philip E. Bourne	Modern Proteomes Contain Imprints of Ancient Shifts in Environmental Chemistry
<a href="#"><u>83</u></a>		Page 12	Wah Chiu, Matthew L Baker	Modeling Protein Components of Biological Nano-Machines in Subnanometer Resolution CryoEM Density Maps
<a href="#"><u>86</u></a>		Page 13	Jeffery G Saven	Theoretical Methods for the Design and Engineering of Proteins
<a href="#"><u>87</u></a>		Page 13	Janet M. Thornton, Gemma L. Holliday, Alex Gutteridge, James Torrance, Gail J. Bartlett, Daniel E. Almonacid, Noel M.O'Boyle, Peter Murray-Rust, John B. O. Mitchell	Using the Three Dimensional Structures of Enzymes to Understand Catalysis
<b>Speaker Abstracts</b>				
<a href="#"><u>8</u></a>		Page 14	Melissa R Landon, Spencer C Thiel, David R. Lancia Jr., Sandor Vajda	Identification of Druggable “Hot Spots” in Ligand Binding Pockets by Computational Solvent Mapping of Proteins
<a href="#"><u>10</u></a>		Page 14	Stefano Lise, David T Jones, Alice Walker-Taylor	Docking protein domains in contact space
<a href="#"><u>15</u></a>		Page 15	Anna CV Johansson, Erik Lindahl	Simulation of Amino Acid Solvation Structure in Transmembrane Helices and Implications for Membrane Protein Folding
<a href="#"><u>17</u></a>		Page 16	Diana K Ekman, Åsa K Björklund, Arne Elofsson	Evolutionary History of Multi-Domain Proteins

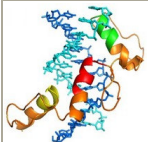
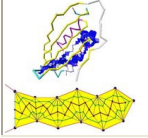
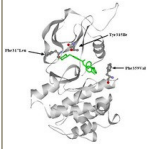
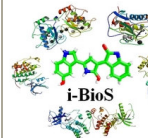

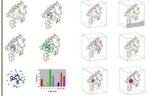
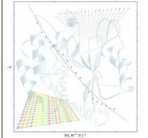
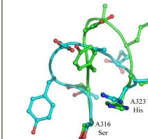
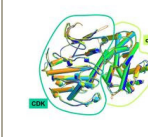
<a href="#">19</a>		Page 17	Michal J Gajda, Marta Kaczor, Janusz M Bujnicki, Anastasia Bakulina, Andrzej Kasperski, Alan M. Friedman, Chris Bailey-Kellogg	FILTREST3D: a Simple Method for Discrimination of Multiple Structural Models Against Spatial Restraints
<a href="#">26</a>	 Fig. 1. Absolute value of Spearman's rank correlation of 3R/3P's deviation of residues in the models compared with the native structure.	Page 17	Marcin Pawlowski, Janusz Bujnicki, Ryszard Matlak	Meta-MQAP: an SVM-based meta-server for the quality assessment of protein models
<a href="#">30</a>		Page 18	Antonio del Sol, Marcos J. Arauzo-Bravo, Dolors Amoros, Ruth Nussinov	Network Robustness and Modularity of Protein Structures in the Identification of Key Residues for the Allosteric Communications
<a href="#">37</a>	 Eqn 1 $C^2 = \frac{1}{N(N-1)} \sum_{i,j} C_{ij}^2$	Page 19	Avner Schlessinger, Marco Punta, Burkhard Rost	Identifying Natively Unstructured Proteins Using Residue Contact Predictions
<a href="#">43</a>		Page 20	Francisco Melo, Alejandro Panjkovich, Andrej Sali, Marc A. Marti-Renom,	Structure Specific Statistical Potentials for Protein Structure Prediction
<a href="#">44</a>		Page 20	Bingding Huang, Michael Schroeder	Integrate the Degree of Burial and Conservation of Surface Residues into Protein-protein Docking
<a href="#">45</a>		Page 21	Sandra Maguid, Gustavo Parisi, Sebastian Fernandez-Alberti, Julian Echave	Sequence-structure-flexibility relationship in protein evolution.
<a href="#">55</a>		Page 22	Iris Antes, Thomas Lengauer	DynaDock: Inhibitor based Refinement of Homology Modeled Protein Structures for Molecular Docking
<a href="#">59</a>		Page 23	Maria S Fornasari, Gustavo Parisi	Protein Structure Diversity and Sequence Divergence in Evolution
<a href="#">62</a>		Page 23	Ingo Ruczinski, Kevin W Plaxco, Tobin R. Sosnick	On the Precision of Experimentally Determined Protein Folding Rates and $\Phi$ Values
<a href="#">63</a>		Page 25	Elon Portugaly, Nathan Linial, Michal Linial	Assessment of Protein Domain Classifications: An Automatic Sequence-based Method and Methods Based on 3D Structures



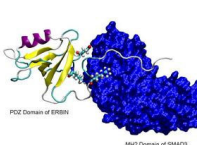
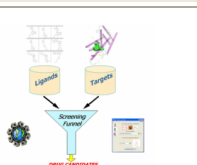
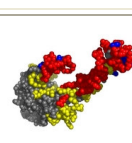


<a href="#"><u>66</u></a>		Page 26	Thomas Funkhouser, Roman Laskowski, Janet Thornton	Matching Volumetric Models of Protein Active Sites
<a href="#"><u>69</u></a>		Page 27	Frank P DiMaio, Jude W Shavlik, George N. Phillips	A Probabilistic Approach to Protein Backbone Tracing in Electron Density Maps
<a href="#"><u>76</u></a>		Page 27	Ilan Samish, Oksana Kerner, David Kaftan, Eran Goldberg, Avigdor Scherz	Datamining the Fourth Dimension from Crystal Structures: Structure-Function-Entropy Relationships in Membrane Proteins
<a href="#"><u>82</u></a>		Page 28	Rafael J Najmanovich, Richard J Morris, Janet M. Thornton	Single-Family Analysis of Binding Site Structural Similarities
<a href="#"><u>84</u></a>		Page 29	Zsuzsanna Dosztányi, Márk Sándor, István Simon	Prediction of Order and Disorder at the Domain Level
<b>Laptop Session Abstracts</b>				
<a href="#"><u>11</u></a>		Page 30	Jorge H Fernandez, Solange M Serrano, Wilson Savino, Ana Tereza Vasconcelos	Bothrostatin, a new RGD-type disintegrin from Bothrops jararaca venom: binding specificity in bothrostatin- $\alpha$ Ib $\beta$ 3/ $\alpha$ v $\beta$ 3 integrin complexes.
<a href="#"><u>12</u></a>		Page 31	Lie Li, Meena Sakharkar, Pandjassaram Kanguene, Jacob Gan, Pandjassaram Kanguene	Structural features differentiate the mechanisms between 2S (2 state) and 3S (3 state) folding of homodimers.
<a href="#"><u>13</u></a>		Page 32	Francisco Torrens, Gloria Castellano	Binding of Water-Soluble, Globular Proteins to Anionic Model Membranes
<a href="#"><u>14</u></a>		Page 33	Karsten M Borgwardt, SVN Vishwanathan, Nicol N. Schraudolph, Hans-Peter Kriegel	Fast Graph Kernels for Proteomics
<a href="#"><u>16</u></a>		Page 34	Talapady N Bhat, Anh D Thi Nguyen, Alexander Wlodawer, Mohamed Nasr	Semantic Web and Chemical Ontology for Fragment-Based Drug Design and AIDS



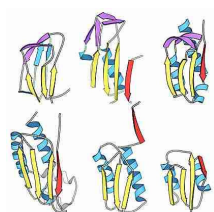
<a href="#"><u>16</u></a>		Page 34	Garima Goel, zaharul ahsan, Srinivasa Reddy Gurram	Development of an effective model for in-silico prediction of distribution parameters of ADMET for viable drug molecules.
<a href="#"><u>18</u></a>		Page 35	Alexander A Kantardjiev, Boris P Atanasov	Electrostatics in Protein Docking: pH-Dependent self-consistent orientation guide
<a href="#"><u>20</u></a>		Page 36	Sumeet Dua, Praveen C Kidambi	Protein Structural Classification using Associative Discrimination of Orthonormal coefficients
<a href="#"><u>21</u></a>		Page 37	Shula Shazman, Yael Mandel-Gutfreund	A Machine Learning Approach for Predicting RNA-binding Proteins from their Three Dimensional Structure
<a href="#"><u>23</u></a>		Page 38	Oliver Sander, Tobias Sing, Francisco S. Domingues, Ingolf Sommer, Andrew Low, Peter K. Cheung, P. Richard Harrigan, Thomas Lengauer	Structural Descriptors Improve the Prediction of HIV Coreceptor Usage
<a href="#"><u>25</u></a>		Page 39	Vishal Kapoor, Sree Nivas gurram Reddy	<i>In Silico</i> Pathway of drug absorption, distribution and Metabolism(ADMET)
<a href="#"><u>27</u></a>		Page 40	Michal Milo, Amiram Goldblum	A Myristoyl Binding Site in Protein Kinases
<a href="#"><u>28</u></a>		Page 40	Claudia Bertonati, Marco Punta, Burkhard Rost, Barry Honig	New Perspectives from Structural Genomics - a Case Study: 5 Recently-Solved North East Structural Genomics Targets Provide New Insights into the Structure and Function of the ASCH-PUA Fold.
<a href="#"><u>31</u></a>		Page 41	Peter Røgen, Per W Karlsson	Quantifying the Relative Positions of Proteins Secondary Structure Elements
<a href="#"><u>34</u></a>		Page 41	Scooter Willis	Predicting co-evolving pairs in Pfam based on Mutual Information of mutation events measured along the phylogenetic tree

<a href="#"><u>39</u></a>		Page 43	Jeff D Sander, Peter C Zaback, Drena L Dobbs, Fengli Fu, Daniel F. Voytas	Designing C2H2 Zinc Finger Proteins to Target Specific Genomic Sites
<a href="#"><u>40</u></a>		Page 43	Bala Krishnamoorthy	Characterization of Protein Structure using Geometry and Topology
<a href="#"><u>41</u></a>		Page 44	Sergio A Alencar, Julio C Lopes, Andrelly M. José	Chemoinformatics approach to Differential Drug Response of the Tyrosine Kinase Domain BCR-ABL to Imatinib in Chronic Myeloid Leukemia Patients
<a href="#"><u>42</u></a>		Page 45	Julio CD Lopes	i-BioS: A System for Inverse Biological Virtual Screening Based on Inverse Docking Approach
<a href="#"><u>47</u></a>		Page 46	Prashanth k. Aitha, Zaharul Ahsan, Srinivasa Reddy Gurram	Toxicity Predication from Chemical Structure as a process of Lead optimization.
<a href="#"><u>54</u></a>		Page 47	Francisco Melo, Evandro Ferrada	Implicit Statistical Potentials for the Characterization and Prediction of Protein-ligand Binding Sites
<a href="#"><u>60</u></a>		Page 47	Oxana V Galzitskaya, Michail Yu Lobanov, Sergiy O Garbuzynskiy	Prediction of Amyloidogenic and Unfolded Regions in Protein Chains
<a href="#"><u>61</u></a>		Page 47	youssef wazady	Conformational Analysis of N-Acetyl-N' Methylcyclopentyl Amide
<a href="#"><u>65</u></a>		Page 48	Mindaugas Margelevicius, Ceslovas Venclovas	A New Approach for Estimation of Statistical Significance in Sequence Profile-Profile Comparisons
<a href="#"><u>68</u></a>		Page 48	Fabrice David, Anne-Lise Veuthey, Yum Lina Yip	Incorporation of 3D Structure Information Into Swiss-Prot Annotation Workflow
<a href="#"><u>73</u></a>		Page 49	Henry van den Bedem, Ankur Dhanik, Jean-Claude Latombe, Ashley M. Deacon	Protein Dynamics from X-Ray Crystallography Data: Modeling Structural Heterogeneity with Inverse Kinematics
<a href="#"><u>74</u></a>		Page 50	Xueping Quan, Peter Doerner, Dietlind L Gerloff	Partner Prediction for Transient Protein Interactions within sets of Paralogues: CDK-Cyclin homologue complexes in <i>A.thaliana</i>

<a href="#"><u>75</u></a>		Page 51	Bruno Contreras-Moreira, Julio Collado-Vides	Comparative footprinting of DNA-binding proteins
<a href="#"><u>78</u></a>		Page 51	Nebojsa Jojic, Manuel J. Reyes-Gomez, David Heckerman, Carl Kadie, Ora Schueler-Furman	Learning MHC I - Peptide Binding
<a href="#"><u>79</u></a>		Page 52	Chavent Matthieu, Claire Nourry, Nadine Deliot, Jean P. Borg, Bernard Maigret	Use of Docking Tools and Molecular Dynamic Simulations to Predict the Interaction Between the MH2 domain of SMAD3 and the PDZ Domain of ERBIN
<a href="#"><u>80</u></a>		Page 52	Alexandre Beautrait, Bernard Maigret	VSM-G: the Virtual Screening Manager Platform for Computational Grids. Use for the Identification of Putative Ligands of the Liver X Receptor.
<a href="#"><u>81</u></a>		Page 53	Michael Terribilini, Jeffry Sander, Byron Olson, Jae-Hyung Lee, Robert Jernigan, Vasant Honavar, Drena Dobbs	A Computational Method to Identify Amino Acid Residues Involved in Protein-RNA Interactions

## Invited Speaker Abstracts

**72 Nick V Grishin**<sup>1</sup> <sup>1</sup>U. of Texas Southwestern Medical Center & HHMI, grishin@chop.swmed.edu



### **Dramatic Evolutionary Changes in Protein Structures**

**Short abstract:** Most common molecular events in sequence evolution are point mutations, insertions/deletions and non-homologous recombination. Effects of these events on protein spatial structure will be discussed, and a few examples of significant structural rearrangements that frequently lead to fold change in evolution will be shown.

**77 Christopher Dupont**<sup>1</sup>, **Song Yang**<sup>2</sup>, **Brian Palenik**<sup>3</sup>, **Philip E. Bourne**<sup>4</sup>

<sup>1</sup>Scripps Institute of Oceanography, U. of California San Diego, cdupont@ucsd.edu <sup>2</sup>Dept. of Chemistry and Biochemistry, U. of California San Diego, soyang@sdsc.edu <sup>3</sup>Scripps Institute of Oceanography, U. of California San Diego, bpalenik@ucsd.edu, <sup>4</sup>Dept of Pharmacology, U. of California San Diego, bourne@sdsc.edu

### **Modern Proteomes Contain Imprints of Ancient Shifts in Environmental Chemistry**



**Short abstract:** The Gaia hypothesis states that life fosters and maintains suitable conditions for itself by helping to create an environment on Earth suitable for its continuity. Using structural bioinformatics to explore the metallomes or species across the Superkingdoms we provide data supporting the Gaia hypothesis.

**Full abstract:** Due to the advent of oxygenic photosynthesis, global trace metal bioavailability has changed dramatically through time potentially affecting the biological usage of metals. Using structural bioinformatics, a census of the abundance and type of Fe, Zn, Mn, and Co-binding protein structures within the proteomes of Archaea, Bacteria, and Eukarya was conducted. Analysis of the metallomes from 312 species reveals that the overall abundances of these metal-binding structures scale to proteome size as a power law, but with different slopes for each Superkingdom. The scaling differences between the Prokaryotes and Eukaryotes are consistent with the evolution of Eukarya in an oxic environment, a hypothesis further supported by the observed modes of Fe binding. These fundamental stoichiometries

imply a closed feedback loop, where the emergence of Eukarya occurs after oxygen-induced changes in trace metal geochemistry.

**83 Wah Chiu**<sup>1</sup>, **Matthew L Baker**<sup>2</sup> <sup>1</sup>MCMI, Baylor College of Medicine, wah@bcm.edu <sup>2</sup>mbaker@bcm.edu

### **Modeling Protein Components of Biological Nano-Machines in Subnanometer Resolution CryoEM Density Maps**



**Full abstract:** Cryo-electron microscopy (cryoEM) is emerging as a structural technique capable of resolving 3-D structures of large biological nano-machines at subnanometer resolutions such as spherical virus particles, ion channels and chaperonins. At this resolution, protein secondary structure elements, including long alpha helices and large beta sheets, can be computationally identified and annotated. However, structural analysis of these large nano-machines are complicated as they are generally made of multiple domains and/or components, each with distinct protein folds. To decipher these structures, we have employed various analysis and modeling approaches utilizing the cryoEM density maps as constraints in the modeling process. When the density is well

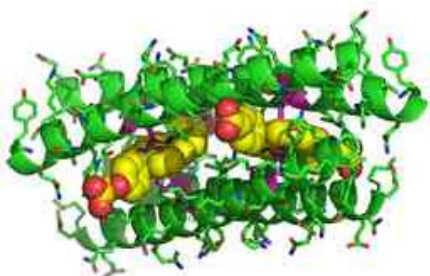
resolved, (e.g. 4-5<sub>Å</sub> resolution) such as GroEL (determined

by S Ludtke and DH Chen, unpublished), the observed alpha helices and beta strands in the cryoEM density can be connected, resulting in a polypeptide backbone trace. However, some of the cryoEM density maps are determined to a lower resolution (e.g. 6-10<sub>Å</sub> resolution in many icosahedral virus particle density maps). In these cases, both constrained homology and ab initio modeling can be utilized. When a related protein structure is known, homology modeling can produce structural models. Constrained by the cryoEM density map, the models can be improved significantly through iterative refinement of sequence alignment and model selection. This approach has been validated through the analysis of the outer capsid protein of the rice dwarf virus, determined first by cryoEM to 6.8<sub>Å</sub> resolution and subsequently using X-ray crystallography. In the absence of any related structural homolog, it is still possible to generate structural models. Structural models can be built directly from the density map, through ab initio methods. Like the aforementioned homology modeling approach, this method may also be constrained by the cryoEM density. An example of this approach has been demonstrated by modeling a small capsid protein from herpes simplex virus type-1 (Vp26, 26 kDa). The resulting model is consistent with existing biochemical and genetic data related to this protein. These structural analysis and modeling methods present a comprehensive approach to maximizing the structural information obtained from intermediate resolution cryoEM density maps.

Acknowledgements: This research has been supported by NCR and NSF. This work has been in collaboration with Tao Ju at Washington U., Maya Topf and Andrej Sali at UCSF, David Baker at the U. of Washington and William Wedemeyer at Michigan State U..

**86 Jeffery G Saven**<sup>1</sup> <sup>1</sup>U. of Pennsylvania, saven@sas.upenn.edu

## Theoretical Methods for the Design and Engineering of Proteins



**Short abstract:** Recent theoretical methods for identifying the properties of amino acid sequences likely to fold to a given three-dimensional structure will be presented, as will several examples of structures so designed that have been experimentally synthesized and characterized.

**Full abstract:** Protein design opens new ways to probe folding, to facilitate the study of proteins, and to engineer new biomolecular systems. Such design is complicated, however, by the conformational complexity of proteins and by the large numbers of possible sequences. Recent theoretical methods for identifying the properties of amino acid sequences likely to fold to a given three-dimensional structure will be presented, as will several examples of structures so designed, which have been experimentally synthesized and

characterized.

**87 Janet M. Thornton**<sup>1</sup>, **Gemma L. Holliday**<sup>2</sup>, **Alex Gutteridge**, **James Torrance**, **Gail J. Bartlett**, **Daniel E. Almonacid**, **Noel M.O'Boyle**, **Peter Murray-Rust**, **John B. O. Mitchell**<sup>1</sup>EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, thornton@ebi.ac.uk <sup>2</sup>European Bioinformatics Institute, gemma@ebi.ac.uk

## Using the Three Dimensional Structures of Enzymes to Understand Catalysis

**Short abstract:** Traditionally each enzyme has been studied individually, but as more enzymes are characterised it is now timely to revisit the molecular basis of catalysis, by comparing different enzymes and their mechanisms, and to consider how complex pathways and networks may have evolved.

**Full abstract:** Enzyme activity is essential for almost all aspects of life. With completely sequenced genomes, the full complement of enzymes in an organism can be defined, and 3D structures have been determined for many enzyme families. Traditionally each enzyme has been studied individually, but as more enzymes are characterised it is now timely to revisit the molecular basis of catalysis, by comparing different enzymes and their mechanisms, and to consider how complex pathways and networks may have evolved.

One aspect of how enzymes perform catalytic reactions, is how do they use the limited set of residue side chains which form their 'catalytic toolkit'? Many catalytic functions are performed by combinations of different residues that form catalytic units that are found repeatedly in different, unrelated enzymes. The majority of catalytic units either provide charged groups, to polarise substrates and stabilise transition states; or modify the pKa values of other residues in order to provide more effective acids and bases. However only a small selection of the different possible catalytic units are commonly used to provide these functions. We show which combinations of residues are used by enzymes, and how their interactions affect their properties. We introduce the concept of the catalytic unit: simple combinations of two or more residues such as the serine protease catalytic triad, which are repeatedly used in similar roles by different, unrelated enzymes. The view of catalysis we present here is undoubtedly a simplification. However, these units provides a useful framework for understanding the chemistry performed by enzymes, and in the long-term may help develop techniques for predicting the mechanisms of enzymes from structure de novo. (Gutteridge & Thornton, 2005) To understand catalysis better we have also developed MACiE (Mechanism, Annotation, and Classification in Enzymes), a database of the chemical mechanisms of enzymatic reactions. The MACiE dataset evolved from that published in the Catalytic Site Atlas (CSA) (Bartlett et al. (2002); Porter et al. (2004)) and each entry is selected so that it fulfils the following criteria. There must be a 3-dimensional crystal structure of the enzyme deposited in the Protein Databank (PDB); the mechanism is relatively well understood mechanism; only one representative of each homologous protein family is included (H level of the CATH code – a hierarchical classification system of protein domain structures (Orengo et al. (1997)) - unless there is a homologue with a significantly different chemical mechanism. Details of each proposed enzyme mechanism are stored in a MySQL database. Research performed using MACiE to reveal which reaction steps are most common will be described. ACKNOWLEDGEMENT: We would like to thank the EPSRC, BBSRC, IBM, Chilean Government and Cambridge Overseas Trust for funding and Unilever for supporting the Centre for Molecular Science Informatics. JBOM and NMOB acknowledge the BBSRC grant BB/C51320X/1 for funding.

## REFERENCES

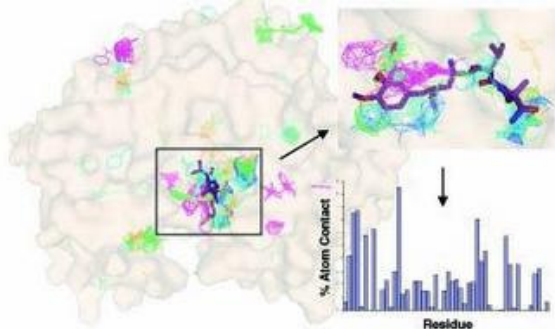
- Gutteridge A & Thornton J.M. (2005) Understanding nature's catalytic toolkit. Trends Biochem Sci. 30(11):622-9.  
Porter,C.T., Bartlett,G.J., Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucl. Acids. Res., 32, D129-D133.  
Holliday, G.L., Bartlett, G.J., Almonacid, D.E., O'Boyle, N.M., Murray-Rust, P., Thornton, J.M. & Mitchell, J.B. (2005) MACiE: a database of enzyme reaction mechanisms. Bioinformatics. PMID 16188925



## Speaker Abstracts

8 Melissa R Landon<sup>1</sup>, Spencer C Thiel<sup>2</sup>, David R. Lancia Jr., Sandor Vajda <sup>1</sup>Boston U., [mlandon@bu.edu](mailto:mlandon@bu.edu)  
<sup>2</sup>[cank@bu.edu](mailto:cank@bu.edu)

### Identification of Druggable “Hot Spots” in Ligand Binding Pockets by Computational Solvent Mapping of Proteins



**Short abstract:** Here we describe the application of the CS-Map algorithm to the identification of druggable subsites within a ligand binding pocket. The accuracy of CS-Map results is illustrated on a set of fifty proteins for which both high affinity drug-like compounds have been developed and structural information is available.

**Full abstract:** Druggability is defined as the capability of a protein to bind lead-like compounds with high affinity. This property has been described previously in NMR-based screening studies that reported a high correlation between the hit-rate and the druggability of a specific binding pocket on a protein<sup>1</sup>. Here we describe the application of a computational method for binding site prediction, the CS-Map algorithm<sup>2</sup>, to

the identification of druggable subsites, or “hot spots”, within a ligand binding pocket. Unlike other existing computational methods for binding site identification, the CS-Map algorithm is capable of resolving the affinities and binding modalities of lead-like fragments on a residue-level within the binding pocket. This is accomplished via a multi-step method that involves (1) placement of a lead-like chemical fragment on the surface of the protein in both predicted and unknown binding locations (2) a search for low energy binding conformations (3) a refined free energy evaluation of the conformations resulting from the search and (4) interaction-based clustering of fragments bound with low free energies. Mapping of multiple fragments on an individual basis results in the creation of consensus sites, e.g. hot spots, established by the aggregation of low energy clusters of multiple fragment types at specific locations on the protein. Further analysis of the consensus sites allows for the characterization of contacts that particular residues in the binding pocket make with chemical fragments. Hydrogen bonding and hydrophobicity patterns can also be extracted from the data, resulting in a pharmacophore depiction of the binding region.

The accuracy of the CS-Map algorithm in the prediction of druggable hot spots is illustrated by the analysis of a set of fifty proteins for which both high affinity drug-like compounds have been developed and structural information is available through the Protein Data Bank (<http://www.pdb.org>). Hot spots resulting from the mapping data for these proteins are consistent with the interactions that the known drugs make with each protein. Of particular interest is Renin aspartic protease, an important therapeutic target for the treatment of hypertension. While several peptide-like Renin inhibitors have been developed<sup>3</sup>, currently only one therapeutic agent, Aliskiren<sup>4</sup>, has entered Phase III clinical trials. Mapping results for Renin are consistent with the binding conformation of Aliskiren versus peptidomimetics, highlighting the importance of particular residue contacts in the peptide binding pocket to the development of high affinity, lead-like compounds for Renin inhibition. Based on these results, we conclude that the information provided by CS-Map can contribute substantially to in silico fragment-based drug discovery efforts.

#### References:

1. Hajduk, P.J., Huth, J.R., Fesik, S.W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* 48 (2005) 2518-25.
2. Silberstein, M., Dennis, S., Brown, L., Kortvelyesi, T., Clodfelter K., Vajda, S. Identification of substrate binding sites in enzymes by computational solvent mapping. *J. Mol. Biol.* 332 (2003) 1095-113.
3. Tong, L., Pav, S., Lamarre, D., Dimoneau, B., Lavallee, P., Jung, G. Crystallographic studies on the binding modes of P2-P3 butanediamide renin inhibitors. *J. Biol. Chem.* 270 (1995) 29520-29524.
4. Wood, J.M., et. al. Structure-based design of aliskiren, a novel orally effective renin inhibitor. *Biochem. Biophys. Res. Commun.* 308 (2003) 698-705.

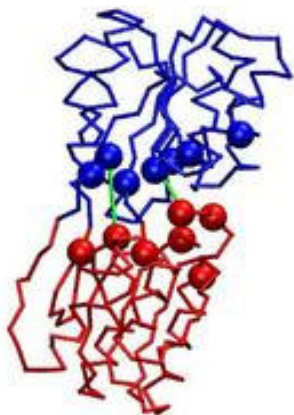
10 Stefano Lise<sup>1</sup>, David T Jones, Alice Walker-Taylor <sup>1</sup>Department of Biochemistry and Molecular Biology, U. College Loondon, [lise@biochem.ucl.ac.uk](mailto:lise@biochem.ucl.ac.uk)

### Docking protein domains in contact space

**Short abstract:** We present a novel method that attempts to dock protein domains using a contact map representation. It is based on a scoring function that combines structural, physicochemical and evolutionary information. The method correctly predicts some contacts across the interface and can complement effectively other computational methods.

**Full abstract:** Many biological processes involve the physical interaction between protein domains. Understanding these functional associations requires knowledge of the molecular structure. Experimental investigations though present considerable difficulties and there is therefore a need for accurate and reliable computational methods.

In this work we present a novel method that seeks to dock protein domains using a contact map representation. Rather than providing a full three dimensional model of the complex, the method predicts contacting residues across the interface. It



works therefore at an intermediate level between binding site predictions and standard docking algorithms. The former methods attempt to identify the interface residues on a protein without specifying the contacts they actually form, the latter aim to provide a detailed atomic model of the putative complex.

We use a scoring function that combines structural, physicochemical and evolutionary information (e.g. local shape complementarity, conservation, interaction potential, etc.). Each potential residue contact is assigned a value according to the scoring function and the hypothesis is that the real configuration of contacts is the one that maximizes the score. The search is performed with a simulated annealing algorithm directly in contact space.

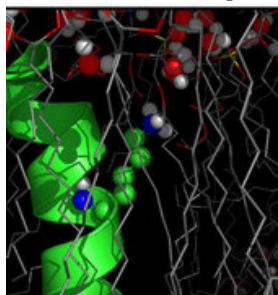
We have tested the method on interacting domain pairs that are part of the same protein (intra-molecular domains). We show that it correctly predicts some contacts and that predicted residues tend to be significantly closer to each other than other pairs of residues in the same domains. Moreover we find that predicted contacts can often discriminate the best model (or the native structure, if present) among a set of optimal solutions generated by a standard docking procedure.

Overall, contact docking appears feasible and able to complement effectively other computational methods for the prediction of protein-protein interactions. With respect to more standard docking algorithms it might be more suitable to handle protein conformational changes and to predict complexes starting from protein models.

**15 Anna CV Johansson<sup>1</sup>, Erik Lindahl<sup>2</sup>** Stockholm Bioinformatics Center, <sup>1</sup>annj@sbc.su.se <sup>2</sup>lindahl@sbc.su.se

## Simulation of Amino Acid Solvation Structure in Transmembrane Helices and Implications for Membrane Protein Folding

**Short abstract:** We report on extensive molecular dynamics simulations used to quantitatively study the atomic scale



structure and dynamics of interactions between hydrophilic residues in transmembrane helices and the bilayer environment. The main conclusion is that hydrophilic residues retain solvation water in amounts that correlate well with experimental hydrophobicity scales.

**Full abstract:** Membrane proteins play key roles in a wide range of processes in the cell, including transport and signaling. The classical view has long been that exposed residues are almost exclusively hydrophobic due to the non-polar lipid membrane environment. There are however transmembrane helices containing charged and polar amino acids and it has been shown experimentally by e.g. Hessa *et al.* [1]

that it is possible to incorporate these hydrophilic amino acids into the membrane as long as they are counterbalanced by enough surrounding hydrophobic residues. Several studies have investigated interactions between membrane proteins and lipid bilayers in the context of

integration of polar and charged amino acids in transmembrane segments and the effects this might have on structure and stability. Polar residues in transmembrane segments tend to be less mutable, indicating their structural and functional importance [2], they have e.g. been shown both to drive association of transmembrane helices in the membrane [3] and bind Heme groups in Cytochrome C oxidase [4]. Studies on existing structures of membrane proteins have shown that the side chains of polar residues located in lipid bilayers tend to be directed away from the membrane core [5, 6], a phenomenon referred to as snorkeling. Snorkeling causes a N to C-terminal composition bias since most polar amino acids are better able to snorkel their polar atoms away from the membrane core at the N-terminus than at the C-terminus [5]. It has also been observed that polar amino acids can form hydrogen bonds with lipid head groups and intramembrane solvation water [7], but to our knowledge the phenomenon has not yet been further evaluated or quantified.

Here, we report on extensive molecular dynamics simulations (3.5 microseconds of aggregated time) used to quantitatively study the atomic scale structure and dynamics of interactions between hydrophilic residues and the bilayer environment. The structural effects of amino acid substitutions in transmembrane helices have been evaluated and quantified using a model system similar to experimental

ones [8, 1]. Results have been classified not only for all amino acids, but also as a function of the z-position in the helix. Polar and charged residues retain significant amounts of solvation water, even to the extent where many of them exist in a microscopic water environment without major distortion of the membrane. Both solvation water amounts and hydrogen bonding patterns exhibit clear position dependence; hydration water around amino acids in the headgroup region exchange rapidly with the bulk solvent, while water retained deeper into the membrane normally requires 5-10 nanoseconds (sometimes even < 15ns) for the exchange to happen. Basic sidechains are found to frequently form hydrogen bonds with lipid carbonyl groups, a pattern we also

confirm by a narrower dip in their relative occurrence inside bilayers compared to acidic residues. Interestingly, this also seems to have caused a slight evolutionary pressure on the overall amino acid distribution in membrane proteins: after subtracting the general positive-inside trend, the basic sidechains exhibit an increased relative occurrence at exactly the same z-coordinates as where the



carbonyl groups are located. Both snorkeling effects as well as N-C terminal composition bias are reproduced by simulations. Further, when pairs of charged residues are introduced close to the center of transmembrane helices, the strong snorkeling preference in combination with their relative locations will cause free helices to tilt (charges on opposite sides) or bend (charges on same side).

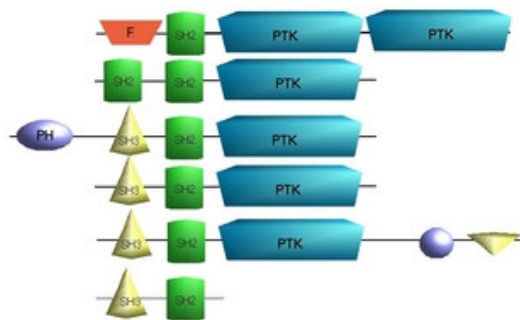
The total amount of hydration water retained around different types of amino acid side chains correlate well both with the experimental Wimley-White (wateroctanol)[9] and the in vivo Hessa/von Heijne hydrophobicity scales [1]. In other words, these results strongly support the hypothesis that even the translocon-mediated membrane protein insertion does not involve any active selection iter, but merely acts as a catalyzer to remove the entropic barrier associated with insertion. In addition, the retained water has quite interesting implications for membrane helix aggregation. Moderately polar individual helices might actually be more water-solvated than previously thought, at least in the sense that there are significant amounts of water present around them. This raises the possibility of a split-timescale aggregation process, where even loose membrane helix contact can lead to a rapid reduction in water-lipid surface (and associated drop in free energy), followed by slower sidechain packing and eventual water expulsion.

#### References

- [1] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, G. von Heijne, Recognition of transmembrane helices by the endoplasmic reticulum translocon., *Nature* 433 (7024) (2005) 37781.
- [2] N. J. Tourasse, W. H. Li, Selective constraints, amino acid composition, and the rate of protein evolution, *Mol Biol Evol* 17 (4) (2000) 65664.
- [3] F. X. Zhou, H. J. Merianos, A. T. Brunger, D. M. Engelman, Polar residues drive association of polyleucine transmembrane helices, *Proc Natl Acad Sci U S A* 98 (5) (2001) 22505.
- [4] S. Iwata, C. Ostermeier, B. Ludwig, H. Michel, Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*, *Nature* 376 (6542)(1995) 6609.
- [5] A. K. Chamberlain, Y. Lee, S. Kim, J. U. Bowie, Snorkeling preferences foster an amino acid composition bias in transmembrane helices., *J Mol Biol* 339 (2)(2004) 4719.
- [6] E. Granseth, G. von Heijne, A. Elofsson, A study of the membrane-water interface region of membrane proteins., *J Mol Biol* 346 (1) (2005) 37785.
- [7] J. A. Freites, D. J. Tobias, G. von Heijne, S. H. White, Interface connections of a transmembrane voltage sensor, *Proc Natl Acad Sci U S A* 102 (42) (2005) 1505964.
- [8] T. Hessa, S. H. White, G. von Heijne, Membrane insertion of a potassium channel voltage sensor, *Science* 307 (5714) (2005) 1427.
- [9] W. C. Wimley, S. H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, *Nat Struct Biol* 3 (10) (1996) 8428.

**17 Diana K Ekman<sup>1</sup>, ,sa K Björklund<sup>2</sup>, Arne Elofsson<sup>1</sup>** <sup>1</sup>Stockholm Bioinformatics Center, Stockholm U., diaek@sbcsu.se <sup>2</sup>asbj@sbcsu.se

#### Evolutionary History of Multi-Domain Proteins



**Short abstract:** A majority of the eukaryotic proteins are multidomain proteins. Here, we present a study of the evolution of new domain architectures through fusion and internal duplication. We found that architectures are created mainly through insertions/deletions of domains at the terminals. Expansion of domain repeats is also an important contribution.

**Full abstract:** The eukaryotic proteomes consist mainly of multi-domain proteins that have been created through fusion, deletion and internal repetition of genes, or parts of genes. The fraction of multi-domain proteins is similar in eukaryotes of different complexity, from yeast to human, and significantly higher than in prokaryotes.

The expansion of protein domain repeats is believed to have been

important particularly for the evolution of higher eukaryotes since large portions of their proteomes consist of repeating protein domains. Here we investigate the creation of new domain architectures through these different evolutionary events. First, a method for identifying the domain architecture from which each protein originates is required. For that we used the domain distance measure, which is defined as the number of domains that differ between two domain architectures. The domain distance relates to the number of domain insertions, deletions and repetitions that are required to transform one architecture to another. Domains that are not detected by the assignment procedure are problematic in studies of domain architectures, since non-existing domain architectures may be found and other existing architectures missed. To reduce this effect we included domains predicted with low confidence if the domain was part of a domain repeat or if it resulted in a domain architecture that was previously identified in another protein. Regions with no detected domains were treated as domains with no homologs, i.e. orphan domains.

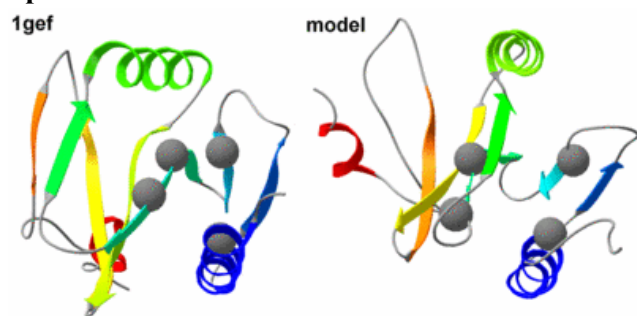
The evolutionary events that distinguish a domain architecture from its most similar architectures were counted. We found that both indels and internal repetitions are frequent, whereas exchanges of domains are rare. Further, indels occur mainly at the N- and C-terminals, while repeated domains may also be inserted between domains. Also in many other respects,

domain repeats behave differently from indels, for example, while indels mainly involve a single domain, repeat expansion frequently involve multiple domains. We found, using internal sequence similarity, that some domain families show distinct duplication patterns, e.g. nebulin domains have mainly been expanded seven domains at a time, while duplications of many other domain families involve varying numbers of domains. However, these duplication patterns show no dependence on the size of the domains. Finally, the rapid expansion of domain repeats may to some extent be explained by exon shuffling.

The evolution of architectures was studied in a number of eukaryotes, including yeast and a plant. Obviously, new domain architectures have evolved in every major evolutionary stage. Considerable expansions of domain architectures have occurred, for example, at the emergence of multicellularity and vertebrates.

**19 Michal J Gajda<sup>1</sup>, Marta Kaczor, Janusz M Bujnicki<sup>2</sup>, Anastasia Bakulina, Andrzej Kasperski, Alan M. Friedman, Chris Bailey-Kellogg** International Institute of Molecular and Cell Biology in Warsaw, Poland, <sup>1</sup>mgajda@genesilico.pl <sup>2</sup>iamb@genesilico.pl

### **FILTREST3D: a Simple Method for Discrimination of Multiple Structural Models Against Spatial Restraints**



#### **Short abstract: FILTREST3D**

<http://filtrest3d.genesilico.pl/> is a new method for discrimination of alternative models against mixed sets of restraints (e.g. spatial distances) derived from experimental analyses as well as from computational predictions (e.g. solvent accessibility) or arbitrarily defined by the user. It can be used to analyze models of individual proteins and complexes.

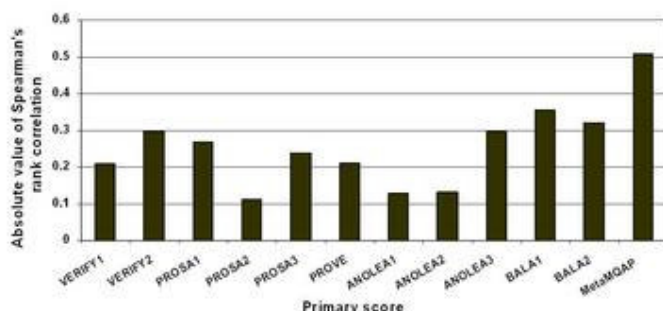
**Full abstract:** Automatic methods for protein structure prediction (fold-recognition and de novo folding methods as well as docking programs) typically produce large sets of alternative models, which often include

approximately correct structures. However, unless the approximately correct models are very close to the native structure (<2.), they are often indistinguishable from completely wrong models based on the energy calculation. On the other hand, such native-like but imperfect models can be often easily identified based on data from experimental analyses that can be encoded as spatial restraints (e.g. approximate distances obtained from cross-linking experiments, surface accessibility from chemical probing, molecule shape from EM or SAXS etc.).

FILTREST3D is a new, simple yet very efficient method for discrimination of a large number of alternative models against mixed sets of restraints derived from experimental analyses as well as from computational predictions (e.g. solvent accessibility, amino acid contact maps) or arbitrarily defined by the user. It can be used to analyze models of individual proteins, protein-protein and protein-nucleic acid complexes. FILTREST3D is available both as a standalone method and as a web server at <http://filtrest3d.genesilico.pl/>.

**26 Marcin Pawlowski<sup>1</sup>, Janusz Bujnicki<sup>2</sup>, Ryszard Matlak** International Institute of Molecular and Cell Biology in Warsaw, <sup>1</sup>marcinp@genesilico.pl <sup>2</sup>iamb@genesilico.pl

### **Meta-MQAP: an SVM-based meta-server for the quality assessment of protein models**



**Fig. 1.** Absolute value of Spearman's rank correlation of MQAPs scores with the deviation of residues in the models compared with the native structures.

(review: (Lazaridis and Karplus, 2000)). So far, the development of MQAPs was focused on the global evaluation of protein structure and most of the existing methods were optimized to discriminate between globally correct and incorrect "decoy" structures rather than detection of correct and incorrect fragments (review: (Gilis, 2004)). Even among MQAPs that are capable of generating independent evaluations for each amino acid in the protein structure, it is usually recommended that scores are averaged over long stretches of residues (e.g. 21 aa in the case of VERIFY3D). Systematic assessment

**Short abstract:** Presented Meta-MQAP uses the results of the five primary MQAPs to predict the absolute deviation (in Angstrom) of C-alpha atoms of all residues in the model from their counterparts in the native structure, without the knowledge of the actual native structure

**Full abstract:** Evaluation of model accuracy is one of the most essential steps in protein structure prediction. The existing methods for quality assessment of protein models (MQAPs) are usually based either on the physical effective energy which can be obtained from fundamental analysis of particles forces or on the pseudo energy derived from known protein structures

experiments, e.g. Critical Assessment of techniques for protein Structure Prediction (CASP) and LiveBench demonstrated that models with correct fold can be confidently recognized, especially by the fold-recognition meta-servers (Bujnicki et al., 2001; Lundstrom et al., 2001). However, comparative models, especially those based on remotely related templates, often exhibit local inaccuracies that are difficult to identify by global evaluation, in particular misthreadings of short regions (5-10 residues) corresponding to shifted alignments within individual secondary structure elements (Tramontano and Morea, 2003; Tress et al., 2005).

We proposed that inaccuracies due to local alignment shifts can be identified and corrected by identification of variable conformations in alternative homology models, comparison of their VERIFY3D scores averaged over only 5 neighboring residues, and construction of hybrid models comprising the best-scoring fragments (Kosinski et al., 2003). Our method (termed the “FRankenstein’s monster approach”) turned out to consistently produce very accurate models, especially if regions with initially poor scores were systematically varied to generate additional models for evaluation (Kosinski et al., 2005). However, detailed inspection of cases where we failed to identify the most native-like local conformation based on the VERIFY3D score revealed a considerable variation of scores even among models with similar structural features. Therefore, we decided to carry out a systematic evaluation of the capability of VERIFY3D and several other popular MQAPs to identify the best method for prediction of local accuracy of protein models. However, as the work progressed, we realized that none of the MQAPs we analyzed was sufficiently accurate and robust, and that they all had different strengths and weaknesses. This in turn prompted us to develop a new “meta-predictor” optimized specifically to detect local errors.

**Results:** Five MQAPs (VERIFY3D, PROSA, BALA, ANOLEA, PROVE) were tested on over 3000 approximately correct models from the latest CASP-6 modeling assessment experiment. We found that the absolute values of correlation coefficients between the scores returned by these methods and the deviation of individual amino acids were rather modest. Thus, we developed a meta-predictor based on a support vector machine (SVM) approach. Our Meta-MQAP uses the results of the five primary MQAPs to predict the absolute deviation (in Angstrom) of C-alpha atoms of all residues in the model from their counterparts in the native structure, without the knowledge of the actual native structure. Scores returned by our Meta-MQAP predict much better the local accuracy of models than the scores returned by the best of the ‘primary’ MQAPs.

**Availability:** We implemented the Meta-MQAP as a web server available for free use by all academic users at the URL <https://genesilico.pl/toolkit/>.

**30 Antonio del Sol<sup>1</sup>, Marcos J. Arauzo-Bravo<sup>2</sup>, Dolores Amoros, Ruth Nussinov** Fujirebio Inc., <sup>1</sup>ao-mesa@fujirebio.co.jp <sup>2</sup>ms-arauzo@fujirebio.co.jp

## Network Robustness and Modularity of Protein Structures in the Identification of Key Residues for the Allosteric Communications

**Short abstract:** We represent protein structures as residue interacting networks in the analysis of several protein families. We identify fold centrally-conserved residues, playing key roles in the allosteric communications. The modular decomposition of these protein networks characterizes different allosteric sites, with the fold centrally-conserved residues participating in the inter-modular interactions.

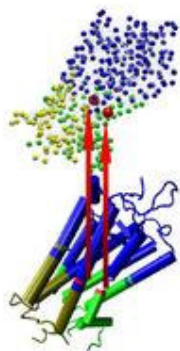
**Full abstract:** Protein structures have evolved towards a robust design, tolerating mutations and environmental changes. At the same time, they are vulnerable to perturbations at key positions or to drastic changes in the environment. This robustness is expected to be reflected in the protein topology. Yet, if we think of protein structures as information processing networks, where the information can be transmitted in a physical (or chemical) form, it would be reasonable to assume that mutations of amino acids crucial for network communications could impair function. It is conceivable that residues, which are presumed to receive and propagate the information, should be lying on the

shortest pathways between most residue pairs in the protein.

**Allostery, Network Robustness and Connectivity:** Allosteric communication is an example of propagation of information transmitting signals from one functional site to another. Here, we represent protein structures as networks of interacting residues, where amino acids are the vertices and their contacts the edges. Our network model of protein structures resembles a robust communication system, where the removal of most of the nodes, with their corresponding edges, does not affect significantly the network’s interconnectedness as described by the characteristic path length. However, when those residues making the most important contribution to generating the small-world character of the network are computationally removed (including their links), the interconnectedness is remarkably affected by a statistically significant increase in the characteristic path length (we termed these residues as “interconnectivity determinants”, or ICD). We found that random rewiring of the edges of the protein networks led to more homogeneous distribution of the residue contribution to the characteristic path length, showing that the communications are no longer maintained by just a few key residues. This indicates that these small-world networks have lapsed into randomness.

**Conserved Interconnectivity Determinants Identify Key Residues for the Allosteric Communications**

We carried out a detailed analysis of seven allosteric protein families (myoglobins, G-protein-coupled receptors, the trypsin class of serine proteases, hemoglobins, oligosaccharide phosphorylases, nuclear receptor ligand-binding domains and retroviral proteases). The family structural alignments identified positions corresponding to the interconnectivity



determinants in the structures of most family members (we termed these residues as “conserved interconnectivity determinants,” or CICD residues) (1). We examined whether CICD residues are related to residues with experimentally demonstrated roles in signal transmission in the seven families. Our results revealed a general correspondence between many of these positions and key residues in allosteric communication. Interestingly, some of the CICD residues in four of the analyzed examples (G-protein-coupled receptors, the trypsin class of serine proteases, hemoglobins, and nuclear receptor ligand-binding domains) were found to be amino acids involved in the networks of statistically coupled residues as predicted by Ranganathan and collaborators (2). We note that here it is not our intention to find networks of important residues possibly involved in allosteric communication. Rather, we show that CICD residues, i.e., centrally conserved residues crucial for maintaining shorter path lengths in the protein network, mediate the signaling process in protein families. The myoglobin family is a particularly interesting example in our analysis. Recent experiments show that this oxygen binding protein is an allosteric enzyme that participates in the catalysis of small molecules. All the CICD residues predicted in this case were identified as residues involved in the myoglobin functions. The HIV-1 protease further constitutes an example where new insights might be gained from an analysis such as the one presented here. Our study detected two CICD residues, which are likely to be involved in the communications between some non-active site residues and the active site. Mutations of these non-active site residues were reported to confer drug resistance on the HIV-1 protease even though they are away from the active site. Further experiments are required to test our predictions. Allosteric regulation is a dynamic process, which implies equilibrium between the active and inactive conformational states. To get an insight into allostery in terms of network communications, we compared between the inactive and active conformations of hemoglobin and of the nitrogen regulatory protein C (NtrC). Our analysis showed that structural changes between the active and inactive conformations lead to a rearrangement of the CICD residues in the two states. This underscores the fact that network communication is dynamic, with altered preferred routes and key residues in different conformational states. Alternate network communications in different regulatory states are advantageous, probably leading to higher efficiency and better control of the transmission of the information. Since these key positions, which are crucial for maintaining the short paths, are centrally-conserved in the protein fold (i.e., are a conserved topological characteristic of the fold rather than being conserved in sequence), it further suggests that it is not necessarily specific residue interactions that are important for regulation. Rather, it is the network characteristics, making the system less sensitive to mutations. In particular, this property of the multiplicity of pathways in the ensembles of different regulatory states confers robustness on the system.

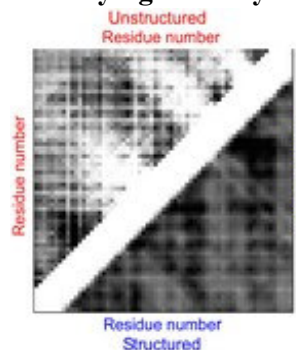
**Network Modularity and CICD Residues:** In an attempt to understand better the role of the CICD residues in the interconnections between different protein regions, we studied the modular decomposition of the analyzed protein networks based on the edge-betweenness clustering algorithm introduced by Girvan and Newman (3). We separated these networks into highly connected groups of amino acid residues (modules) by removing those edges (residue contacts), which lie in the shortest paths between most pairs of vertices. Thus, all shortest paths between modules must go along one of these edges. Results revealed that modules tend to characterize different functional sites (active and regulatory sites) involved in the protein allosteric communications. The majority of CICD residues participate in the inter-modular interactions, supporting the idea that these amino acids play a key role in the transmission of the information between functional regions.

References:

1. del Sol A, Fujihashi H, Amoros D, Nussinov R (2006) Molecular Systems Biology (in press).
2. Suel G, Lockless SW, Wall M, Ranganathan R (2003) Nature Structural Biology 10, 59-69.
3. Girvan M, Newman MEJ (2002) Proc. Natl. Acad. Sci. 99, 7821-7826.

**37 Avner Schlessinger<sup>1</sup>, Marco Punta<sup>2</sup>, Burkhard Rost** Columbia U., <sup>1</sup>as2067@columbia.edu <sup>2</sup>punta@rostlab.org

## Identifying Natively Unstructured Proteins Using Residue Contact Predictions



Eqn. 1 
$$e_i^p = \frac{1}{2L+1} \sum_{a=i-L}^{i+L} c_a^p \cdot M_a$$

**Short abstract:** Recently, increasing attention has been drawn to the study of ‘natively unstructured’ proteins. Here, we combine a position specific contact propensity predictor with an average pairwise energy-like matrix to estimate the energetic contribution of each residue of the protein sequence to the protein foldability. Our novel approach outperforms competing methods.

**Full abstract:** In the past few years, several studies have shown that the ability of a protein to adopt a unique native 3D structure only in specific biochemical or cellular conditions can be crucial for function [1]. Proteins that do not have a completely ordered native structure in physiological conditions are often referred as ‘natively unstructured’ proteins. A typical case is that of a protein that is unfolded when isolated but adopts a unique structure upon binding to a target thereby performing its biochemical function. Therefore, the binding enthalpy of the new interactions compensates for the entropic cost [1].

Recently, the seemingly low propensity of these proteins to form internal residue contacts has been used to develop prediction tools that aim to discriminate natively folded from natively unstructured proteins [2, 3]. However, these methods use sequence features only and cannot accurately capture the effect of long range interactions on the protein structure.



Here, we use PROFcon, a neural network-based inter-residue contact predictor [4], to derive protein-specific contact propensity maps (see figure) and combine these maps with an average pairwise energy-like matrix to predict the presence of unstructured regions in a protein chain. We estimate the energetic contribution of each residue  $i$  to the foldability of a protein  $P$  ( $e_i^P$ ), according to Eqn. 1, where  $c_{ik}^P$  is the protein-dependent contact propensity for the pair  $(i,k)$  and  $M_{ik}$  is its estimated interaction energy [3];  $L$  is a free parameter that can be optimized. We calculate  $e_i^P$  for every residue in a chain  $P$  and smooth the values of the obtained profile by using a window of size  $S$ . Disordered regions can be identified as peaks in this profile. In practice, we define a threshold  $T$  above which a residue is labeled as disordered. In order to estimate the free parameters of our method (i.e.  $L$  in Eqn. 1,  $S$  and  $T$ ) we perform cross-validation on a database of 224 ordered (part of PROFcon test set) and 313 disordered (DisProt dataset [5]) proteins. The performance of our method compares favorably with state-of-the-art methods predicting long disordered regions.

[1] Dyson, H.J. and P.E. Wright, Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 2005. 6(3): p. 197-208.

[2] Garbuzynskiy, S.O., M.Y. Lobanov, and O.V. Galzitskaya, To be folded or to be unfolded? *Protein Sci*, 2004. 13(11): p. 2871-7.

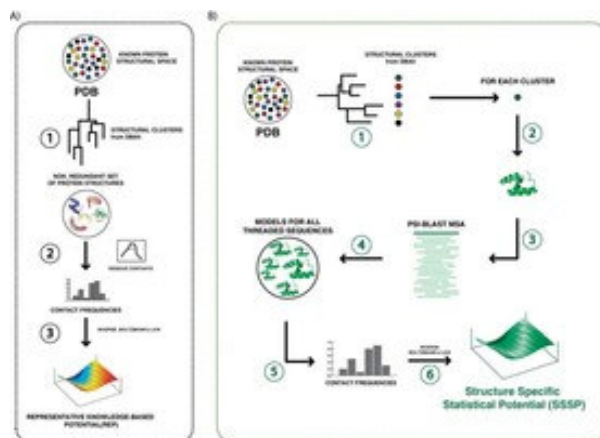
[3] Dosztanyi, Z., et al., The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, 2005. 347(4): p. 827-39.

[4] Punta, M. and B. Rost, PROFcon: novel prediction of long-range contacts. *Bioinformatics*, 2005. 21(13): p. 2960-8.

[5] Vucetic, S., et al., DisProt: a database of protein disorder. *Bioinformatics*, 2005. 21(1): p. 137-40.

**43 Francisco Melo<sup>1</sup>, Alejandro Panjkovich<sup>2</sup>, Andrej Sali, Marc A. Marti-Renom,** Catholic U. of Chile, <sup>1</sup>fmelo@bio.puc.cl <sup>2</sup>sasha.panjkovich@gmail.com

### Structure Specific Statistical Potentials for Protein Structure Prediction



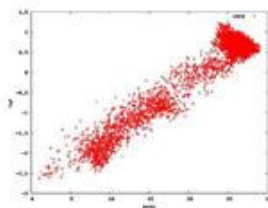
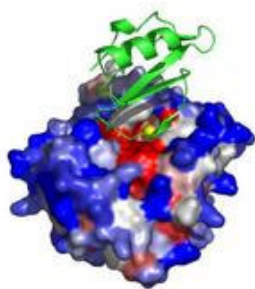
**Short abstract:** We describe a novel approach to derive improved statistical potentials for protein structure prediction, which are specific of a particular fold or structure. The specific potentials take advantage of the known evolutionary record by incorporating homologue sequences that give useful information about kinetics, function and thermodynamics governing protein folding.

**Full abstract:** We will describe a new method to derive statistical potentials for protein structure prediction. This method attempts to incorporate not only structural information from non-redundant native proteins that account mostly for the thermodynamic contributions observed in protein folding, but also takes advantage of evolution records through known homologue protein sequences to incorporate the information of those residues that are either essential for function, kinetically important or kinetically restrictive in the

protein folding process. By this procedure, a specific statistical potential was derived for each known non-redundant protein structure from the PDB, based on a single representative experimental structure and all its known protein sequence homologues. A total of 22,000 structure-specific statistical potentials were derived. Based on a large benchmark of about 10,000 protein structure models built by comparative modeling, we demonstrate a significantly improved performance of the structure specific statistical potentials for the task of fold assessment, as compared to that obtained when using a single representative statistical potential derived from a set of non-redundant protein structures.

**44 Bingding Huang<sup>1</sup>, Michael Schroeder<sup>2</sup>** Biotec, Technical U. Dresden, Germany, <sup>1</sup>bhuang@biotec.tu-dresden.de <sup>2</sup>ms@biotec.tu-dresden.de

### Integrate the Degree of Burial and Conservation of Surface Residues into Protein-protein Docking



**Short abstract:** The binding sites of enzyme-inhibitor complexes usually involve a very deep buried pocket and are more conserved than the rest of surface. In our docking approach BDOCK, the degree of burial and conservation of surface residues are taken into account, which significantly improves the performance of traditional FFT docking method.

**Full abstract:** Protein-protein interactions are fundamental as many proteins mediate their biological function through protein interactions. Most processes in the living cell requires molecular recognition and formation of complexes, which may be stable or transient assemblies of two or more

molecules with one molecule acting on the other, or promoting intra- and intercellular communication, or permanent oligomeric ensembles. The rapid accumulation of data on protein-protein interactions, sequences, structures calls for the development of computational methods for protein docking. Typically docking methods are investigated which attempt to predict the complex structures given the structures of components. Over the past 20 years there have been many computational approaches to dock proteins. These approaches are mostly based on the shape complementarity of structures and the physio-chemical properties of the interfaces. However, these docking approaches are far from perfect and there still remains room for improvement.

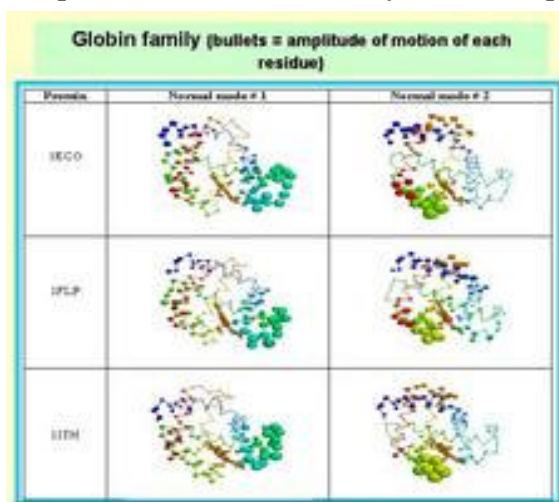
In blind protein-protein docking approaches, it is of great importance that the binding sites are predicted correctly in the first step. Knowing where the binding sites are located on the protein surface can limit the conformational search space and reduce computational time. In the last 10 years, there have been many efforts to predict the protein-protein interaction binding sites based on the analysis of the protein surface properties. However, only a few approaches integrate the interface prediction into docking. The binding sites of enzyme-inhibitor complexes always involve a very deep buried pocket. It is also believed that the binding sites are more conserved than the rest of surface. In this work, we develop a system called BDOCK, which uses both the degree of burial and conservation to improve docking. First, we propose a novel shape complementary scoring function for the initial docking stage, which takes the degree of burial of surface residues into account as different weights. Secondly, in the filter stage, the highly conserved surface residues located around a pocket are considered to be predicted interface residues, which are then used to calculate the tightness of fit as a scoring function to filter the docking solutions generated in the first step.

Our approach is evaluated on the unbound structures of 22 enzyme-inhibitor complexes from Chen benchmark data set. An interface prediction is defined as success if more than 50% of the predicted residues are real interface residues. Here we define a docking solution as near-native structure (hit) if the RMSD between it and the native complex is below 4.5 angstrom. In the first step, the top 2000 solutions based shape complementarity are kept for each complex. The Z-score of the tightness of fit is calculated and the threshold for filtering is set to -1.0. As results, the interface residues are correctly predicted for 19 complexes. Furthermore, 76% success rate is achieved on a large data set of 102 enzyme-inhibitor complexes derived from the PDB database and SCOP database. In the initial stage of docking, our approach can generate some near-native complex structures for all 22 cases, ranging from 1 to 931. In the filter step, the tightness of fit scoring function based on the predicted interface residues improves the docking results by factors of 4-61. Moreover, the tightness of fit significantly improves the ranking of the best hit to the top 10 for 12 cases.

BDOCK is implemented using the BALL library in C++, taking an object-oriented and generic programming approach, which is very easy to extend and adapt new docking algorithms. The source code of BDOCK is available at <http://www.biotech.tu-dresden.de/~bhuang/bdock>.

**45 Sandra Maguid<sup>1</sup>, Gustavo Parisi<sup>2</sup>, Sebastian Fernandez-Alberti<sup>3</sup>, Julian Echave** Universidad Nacional de Quilmes, <sup>1</sup>[smaguid@unq.edu.ar](mailto:smaguid@unq.edu.ar) <sup>2</sup>[gustavo@unq.edu.ar](mailto:gustavo@unq.edu.ar) <sup>3</sup>[seba@unq.edu.ar](mailto:seba@unq.edu.ar)

### Sequence-structure-flexibility relationship in protein evolution.



**Short abstract:** We study the evolutionary divergence of protein backbone dynamics by comparing the  $C\alpha$  flexibility profiles for a large dataset of homologous proteins classified into families and superfamilies. Proteins were structurally aligned and for each alignment we calculated sequence similarity, structural dissimilarity and different measures of backbone flexibility similarity.

**Full abstract:** Internal protein dynamics is essential for biological function. During evolution, protein divergence is functionally constrained: properties more relevant for function vary more slowly than less important properties. Thus, if protein dynamics is relevant for function, it should be evolutionary conserved. In contrast with the well studied evolution of protein structure, the evolutionary divergence of protein dynamics has not been addressed systematically before, apart from a few case studies. X-ray diffraction analysis gives information not only on protein structure but also B-factors, which characterize the flexibility that results from protein dynamics. Here, we study the evolutionary

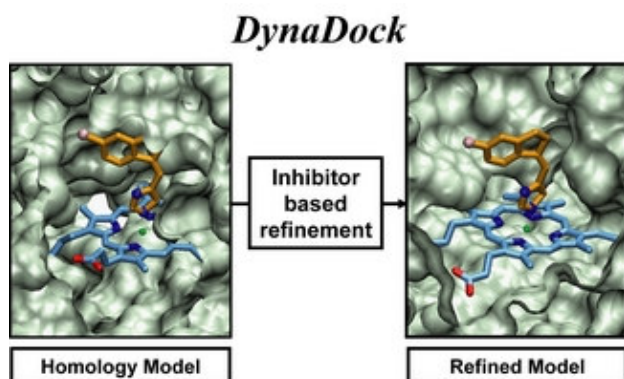
divergence of protein backbone dynamics by comparing the  $C\alpha$  flexibility (B-factor) profiles for a large dataset of homologous proteins classified into families and superfamilies. A dataset of pair alignments of non-homologous protein pairs was used as reference. Proteins were structurally aligned and for each alignment we calculated sequence similarity (Id%), structural dissimilarity (RMSd) and two measures of backbone flexibility similarity: the Spearman rank-order correlation coefficient between the aligned B-factor profiles,  $\rho_B$ , and the Pearson linear correlation coefficient  $r_B$ . Furthermore, the vibrational dynamics of each protein was studied by normal modes analysis (NMA). The low-frequency normal modes describe collective movements that are closely related to the protein's biological function [Berendsen 2000]. The NMA was performed using the Gaussian Network Model (GNM) [Haliloglu 1997] [Bahar 1997]. This model is shown

to be a simple and efficient computational method to study the collective motions of large proteins. In order to explore the common dynamics of homologous proteins, a new procedure was developed [Maguid 2005]. The method allows the comparison of patterns of vibrational motions obtained by GNM and involves the alignment, rearrangement, and Singular Value Decomposition of the normal modes of homologous proteins. We were able to study the common dynamics within a family and superfamily by the identification of collective coordinates that were conserved during the evolution.

We show that  $\text{Ca}$  flexibility profiles diverge slowly, so that they are conserved at family and superfamily levels, even for pairs of proteins with non-significant sequence similarity. We also analyse and discuss the correlations between the divergence of flexibility, sequence, and structure [Maguid 2006]. The present work has gone beyond the comparison of native structures, by comparing backbone flexibility profiles. These contain information on the equilibrium distribution of protein conformations. Thus, the present results imply the conservation not only of the average structure but also of the dispersion of the equilibrium distribution around the average structure. Such distribution is determined by the intramolecular dynamics of the protein. Thus, the observed conservation of flexibility profiles provides indirect evidence of the conservation of protein dynamics. Furthermore, the comparative analysis of the normal modes in homologous proteins reveals different levels of evolutionary conservation at family and superfamily levels for the first 50 modes ordered by increasing frequency values. The differences between the vibrational dynamics of the homologous proteins rapidly increase with the average frequency of their corresponding equivalent modes. The high degree of conservation observed in the few lowest modes indicates that mutations in proteins are constrained to produce limited changes at the tertiary level of structure that guarantee a unique pattern of relative flexibilities that provide protein functionalities. Therefore, we show that mutations within a fold family not only conserve a certain degree of protein structure but also features of their dynamics.

#### References

- Berendsen, J. C., and S. Hayward (2000). *Curr. Opin. Struct. Biol.* 10:165–169.  
 Haliloglu, T., I. Bahar, and B. Erman (1997). *Phys. Rev. Lett.* 79:3090–3093.  
 Bahar, I., A. R. Atilgan, and B. Erman (1997). *Fold. Des.* 2:173–181.  
 Maguid, S., S. Fernandez-Alberti, L. Ferrelli, and J. Echave (2005). *Biophys. J.* 89: 3-13.  
 Maguid, S., S. Fernandez-Alberti, G. Parisi, and J. Echave (2006). *J. Mol. Evol.* (in press).



**55 Iris Antes<sup>1</sup>, Thomas Lengauer<sup>2</sup>** Max-Planck-Institute for Informatics, <sup>1</sup>antes@mpi-inf.mpg.de  
<sup>2</sup>lengauer@mpi-inf.mpg.de

#### **DynaDock: Inhibitor based Refinement of Homology Modeled Protein Structures for Molecular Docking**

**Short abstract:** We present a new refinement strategy and program, DynaDock, for homology modeled protein structures to be used molecular docking. The approach uses an iterative, inhibitor-based refinement protocol and was evaluated on several CYP P450 protein structures and applied during a drug design project searching for inhibitors of CYP11B2.

**Full abstract:** The use of homology modeled structures

for purposes for which detailed information about the protein's interactions is necessary is very often not a straightforward task. The two major reasons for this are: first, the limited accuracy especially of the side chain positions in protein structures obtained by homology modeling and second, the dependency of the structural details for the homology model on the template structures used. Thus often homology modeled structures cannot be used directly for purposes like e.g. molecular docking. An additional refinement and validation of these structures is necessary first. We present a new refinement strategy and program, DynaDock, which performs an iterative, inhibitor-based refinement of homology modeled structures, and discuss the structural improvements gained through our refinement protocol.

Our approach starts with a homology modeled backbone of the target protein, which can be obtained by any homology modeling package. Afterwards side chains are placed with the IRECS module. The resulting homology model is then iteratively refined by alternately docking known inhibitors into the model using FlexX and refining the docked complexes through a special molecular dynamics/Monte Carlo based refinement protocol. If all known inhibitors can be docked with sufficient accuracy and selectivity against non-inhibitors, the corresponding homology model is accepted as the final receptor structure for further docking or virtual screening studies.

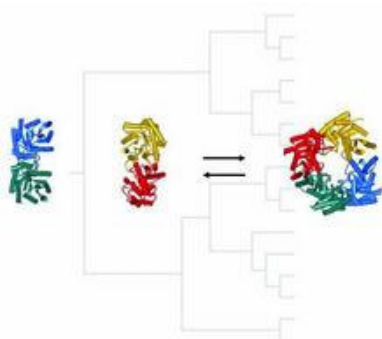
To evaluate our approach we chose several members of the cytochrome P450 protein superfamily. Due to the difficulty of resolving membrane bound proteins, there are very few experimental structures available for mammalian/human CYP P450s. Thus there is an intense ongoing effort to develop homology models for CYP P450 proteins. In addition, our approach was applied and further evaluated in the context of a drug design project in collaboration with the Pharmaceutical Chemistry Department of the Saarland U., searching for selective inhibitors of aldosterone synthase: CYP11B2. CYP11B2 catalyses the final steps of aldosterone synthesis. High aldosterone levels can lead to various cardiovascular diseases and



thus CYP11B2 is an important drug target. Through combined computational and experimental studies several selective and highly potent inhibitors could be found, which are at the moment under further pharmaceutical investigation.

**59 Maria S Fornasari<sup>1</sup>, Gustavo Parisi<sup>2</sup>** Universidad Nacional de Quilmes, <sup>1</sup>silvina@unq.edu.ar <sup>2</sup>gustavo@unq.edu.ar

## Protein Structure Diversity and Sequence Divergence in Evolution



**Short abstract:** Using the Structurally Constrained Protein Evolution model (SCPE), we found that the consideration of the quaternary structure and the different oligomeric states a protein could adopt, enhance the description of the sequence divergence during evolution.

**Full abstract:** A major constraint in protein sequence divergence is the conservation of protein structure. In recent years, there is an emerging picture which describes the native structure of a protein as an ensemble of different conformations. This concept was described earlier in the well known Monod-Wyman-Changeux model[1] for allosteric proteins or the so called pre-existing-equilibrium hypothesis. In this work we study the modelling of protein divergence as a consequence of the occurrence of multiple conformations for a protein native state.

It has been demonstrated that the explicit incorporation of structural information, as protein fold, stability and/or foldability enhances the description of sequence divergence during protein evolution[2;4;5]. To study how protein structure conservation modulates sequence divergence, we developed the Structurally Constrained Protein Evolution model(SCPE, [5]). The SCPE model simulates evolution by introducing random mutations and selecting them for structural conservation. Here, we first analyse SCPE performance when quaternary structural information is incorporated. We found that the consideration of the quaternary structure in the SCPE enhance the model performance when is compared with a monomeric description of the corresponding protein. Furthermore, we also show that the consideration of the different oligomeric states that protein could adopt in our model also enhances the description of its evolution.

For this study we use as test system the protein triosephosphate isomerase from the hyperthermophilic archaeal *Thermoproteus tenax* (TtxTIM)[6]. This protein exists in a dimer/tetramer equilibrium. Although the dimer is inactive, the equilibrium is shifted to the active tetrameric form through a specific interaction with glycerol-1-phosphate dehydrogenase[6].

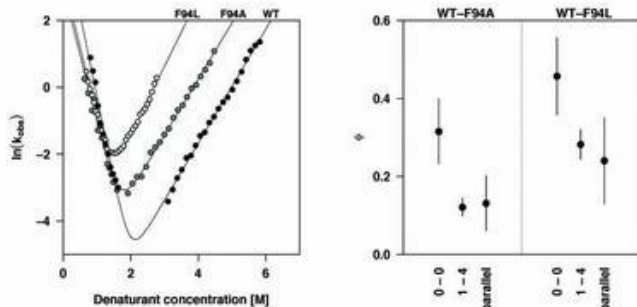
A sequence similarity search using TtxTIM as input was done using BLASTP with default parameters. The retrieved sequences (with E-values below 110-4) were aligned with CLUSTALX. Using this alignment a Neighbor-Joining tree was estimated using maximum likelihood distances calculated with the JTT+F model with gamma rate variation using the program HYPHY[7]. From this tree the subtree containing TtxTIM and close homologous proteins was identified. SCPE simulations were performed with the monomeric, dimeric and tetrameric forms of TtxTIM to obtain a set of site-specific substitution matrices as described previously[3]. Also using HYPHY, with these substitution matrices we calculated the maximum likelihood using the fixed topology (subtree) and the corresponding alignment mentioned above. Likelihood-ratio[7] test was used to study the statistical significance of the different models.

Our results indicate that the best conformation describing the divergence of TtxTIM is the tetrameric form in terms of the maximum likelihood analysis. We observed that the monomer performed worse than the dimer, and the dimer worse than the tetramer. In a profile analysis of the logarithm of the likelihood per position we found that the main difference between the tetramer and the dimer are located at the tetrameric interfaces. It is important to note that TtxTIM does not form stable tetramers but this is the active form of the protein. Then, the tetramer is the conformation subjected to selective pressure to conserve the biological activity, observation also supported by the maximum likelihood calculations.

### References

- [1]Monod, J., Wyman, J., Changeux, JP. 1965. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* 12, 88-118.
- [2]Bastolla, U., Roman, H.E., and Vendruscolo, M. 1999. Neutral evolution of model proteins: diffusion in sequence space overdispersion. *J Theor Biol.* 200:49-64
- [3]Fornasari, M.S., Parisi, G., and Echave, J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol Biol Evol.* 19:352-6
- [4]Govindarajan, S. and Goldstein, R.A. 1997. Evolution of model proteins on a foldability landscape. *Proteins.* 29:461-6
- [5]Parisi, G. and Echave, J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol.* 18:750-6
- [6]Walden, H., Taylor, G., Lilie, H., Knura, T., and Hensel, R. 2004. Triosephosphate isomerase of the hyperthermophile *Thermoproteus tenax*: thermostability is not everything. *Biochem Soc Trans.* 32:305
- [7]Goldman, N., 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36, 182– 198.
- [8]Pond SL, Frost, SD and Muse, SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics.* 21(5):676-9

**62 Ingo Ruczinski<sup>1</sup>, Kevin W Plaxco<sup>2</sup>, Tobin R. Sosnick** <sup>1</sup>Johns Hopkins U., ingo@jhu.edu <sup>2</sup>U. of California at Santa Barbara, kwp@chem.ucsb.edu



## On the Precision of Experimentally Determined Protein Folding Rates and $\Phi$ Values

**Short abstract:**  $\Phi$ -values are the most important experimental benchmark by which theoretical models of protein folding are tested. Addressing the significant controversy that has emerged regarding reliability, we show how to derive valid standard errors for  $\Phi$  estimates, and discuss what determines the precision with which  $\Phi$ -values can be estimated.

**Full abstract:** The protein folding process is an example of a highly specific, spontaneous,

nanometer-scale self-assembly process. It also lies at the heart of an increasing number of disease states, such as Alzheimer's, Parkinson's and Huntington's disease. Lastly, and perhaps most centrally, it is the process by which the information encoded in our genes is expressed in the form of the functional catalysts that promote and control every aspect of cellular life. Thus motivated, the study of protein folding has exploded in the last decade, with literally thousands of biochemical research groups participating in its exploration.

Of particular interest to the community of protein folding aficionados is the structure that the protein adopts in the folding transition state (the state with highest free energy), as that state forms the barrier that ultimately defines the folding pathway. Moreover, there is some evidence that it is this state from which disease-causing aggregation often occurs. Unlike the structure in the initial, unfolded state and the final, folded state however, the structure in the transition state cannot directly be assessed by experimental means such as X-ray crystallography or Nuclear Magnetic Resonance (NMR) Spectroscopy. Therefore, researchers have to rely on different means to assess structure and mechanism in the folding process. The idea underlying this indirect approach is as follows: a mutation is introduced at a locus of interest in the amino acid sequence of the protein. This mutation can have an effect on the change in free energy between the wild type and the mutant folded states, and can also have an effect on the change of free energy in the transition states. If the ratio of change in free energy at the transition state and the change in free energy at the folded state (called the  $\Phi$ -value) is zero, the protein is completely unstructured in the transition state at the locus where the mutation was introduced. If the changes in free energy are the same, i.e. the  $\Phi$ -value is equal to unity, the protein is structured in the transition state at the locus where the mutation was introduced. By introducing mutants at different sequence loci, information can be obtained which parts of a protein are participating in interactions in the transition state, and which are not, yielding critical information about the nature of this defining conformation.

Because  $\Phi$ -value analysis is the only available means of experimentally characterizing the folding transition state it has, in the 15 years since it was first introduced, largely dominated the folding field and is the single most important means by which theories of the folding process are tested experimentally. Thus motivated, most of the major experimental folding laboratories have performed this experiment on dozens of different proteins to date. Recently, however, significant controversy has emerged regarding the reliability with which  $\Phi$ -values can be determined experimentally. Since  $\Phi$  is a ratio of differences between experimental observables, it is extremely sensitive to errors in those observations when the differences are small. Moreover, because both the numerator and denominator in this ratio are defined via kinetic parameters that are all derived from a single experimental data set, correlations between these parameters render error estimation in  $\Phi$  analysis far from straightforward. Perhaps consistent with this, a survey of dozens of key  $\Phi$ -value papers produces *not a single example in which the methods employed for error analysis are described*. Given the extent to which quantitative comparisons between theoretical models of the folding transition state and  $\Phi$ -values have dominated the folding field, the failure of the experimental community to properly calculate confidence intervals for this experimental observable could have dire consequences for the understanding of this key biochemical process.

In this presentation, we address the above issues in detail. In particular, we

- derive an analytical expression for the precision of  $\Phi$  estimates that explicitly takes the dependence between estimates of the changes in free energy of the transition and folded states into account, describe an alternative method that implicitly corrects for the effect, and introduce a [web-based implementation of the above algorithms](#) for general use by the protein folding community;
- discuss the ingredients that determine the precision with which  $\Phi$ -values can be estimated, in particular as a function of  $\Delta\Delta G_N$  (the change of free energy in the folded state between wildtype and mutant);
- determine the precision with which folding rates and  $\Phi$ -values are measured using generally accepted laboratory practices and under conditions typical of our laboratories, by performing blind, replicate measurements in three laboratories, and show that the reproducibility of results derived from those kinetic experiments can compare poorly with the confidence intervals typically reported in the literature,
- explore the effects of other commonly employed techniques of calculating  $\Phi$  from kinetics, such as estimation at non-zero denaturant concentrations and via the assumption of parallel chevron arms, and show that these approaches can

produce significantly different estimates for  $\Phi$ , thus indicating that they are not equivalent measures of transition state structure.

References:

- De Los Rios MA, Muralidhara BK, Wildes D, Sosnick TR, Marqusee S, Wittung-Stafshede P, Plaxco KW, Ruczinski I (2006)

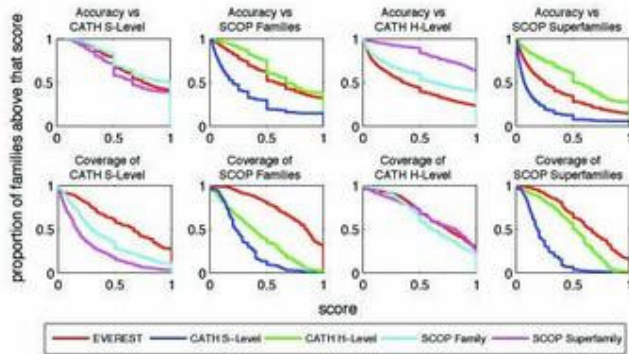
*On the precision of experimentally determined protein folding rates and  $\Phi$ -values*, Protein Science, 15(3): 553-63.

- Ruczinski I, Sosnick TR, Plaxco KW (2006)

*Methods for the accurate estimation of confidence intervals on protein folding  $\Phi$ -values* Protein Sciences (In Publication).

**63** **Elon Portugaly<sup>1</sup>, Nathan Linial<sup>2</sup>, Michal Linial** The Hebrew U. of Jerusalem, School of Computer Science and Engineering, <sup>1</sup>elonp@cs.huji.ac.il <sup>2</sup>nati@cs.huji.ac.il

## Assessment of Protein Domain Classifications: An Automatic Sequence-based Method and Methods Based on 3D Structures



**Short abstract:** EVEREST is an automatic system that identifies and classifies domains within a database of protein sequences. We show that the set of EVEREST families is as similar to the set of CATH families and to the set of SCOP families as the latter two sets are similar to each other.

**Full abstract:** Introduction: EVEREST[6] is an automatic system that identifies and classifies domains within a database of protein sequences. It takes as input a database of protein sequences, and examples of known domain families used for learning the notion of a domain family. Its output is a set of domain families given as automata that are able to identify domains of the family in protein sequences.

SCOP[4] and CATH[5] are two systems that classify domains of known structures. Each of the two systems defines the domain boundaries within experimentally solved structures and classifies the domains into a shallow hierarchy of classes. We present here a comparison of the two structural classifications and of each of them to an exhaustive classification defined by EVEREST on the sequences corresponding to the structures. Differences and similarities between alternative classifications of protein domains have been reported and discussed[3]. Here we show surprising agreement with an automatic sequence-based definition.

**Methods:** Comparing classification systems: Our scheme for comparing two domain classification systems involves the asymmetric evaluation of an *evaluated* family-set with respect to a *reference* family-set. Let  $s(e, r)$  be the score of an evaluated family  $e$  with respect to a reference-family  $r$ , as defined in Evaluating a domain-family. The evaluated family set is tested in two complementary ways:

- **Accuracy:** how many of the evaluated families that intersect with the reference families are good reconstructions of any reference family, as described by the histogram over evaluated families  $e$  of  $s(e) = \max_r s(e, r)$ , where  $r$  runs over the set of reference families. An evaluated family that does not intersect with any reference family is considered new, does not enter the evaluation.
- **Coverage:** how many of the reference families are reconstructed well, as described by the histogram over reference families  $r$  of  $s(r) = \max_e s(e, r)$  where  $e$  runs over the set of evaluated families.

**Evaluating a domain family**

Define  $P(e)$ , the *reference projection* of an evaluated family  $e$  as the set of reference domains that significantly intersect with the domains of the suggested family. A reference domain and an evaluated domain are said to be significantly intersecting if their intersection is at least 80% of the shorter of the two. Evaluated families whose reference projection is empty cannot be evaluated by the reference set, and are ignored.

The score of  $e$  with respect to  $r$  is defined as the size of the intersection of  $P(e)$  with  $r$  divided by the size of their union.

**Databases:** 22728 distinct PDB[1] sequences were downloaded from the PDB on February 2006. SCOP domains were taken from ASTRAL release 1.69[2]. CATH release 2.6.0 was used.

To achieve common ground, both SCOP and CATH domains were first mapped to PDB sequences. Two mappings were performed for each system. An inclusive mapping for use when the system plays the role of the evaluated set, and an exclusive mapping for use when the system plays the role of the reference set. For the exclusive mapping, we have used non-redundant versions of the data, namely ASTRAL with 95% similarity cutoff, and one representative per CATH L-Level group. We also removed all non-continuous domains for the exclusive mapping. In both types of mapping, only exact sequence matches were accepted. The exclusive mapping yielded 14,900 domains for CATH and 11,358 domains for SCOP. EVEREST release 2 contains 20,230 domain families. Of those 14,247 families were found on the PDB sequences.

**Results and Conclusions:**

We use four reference-family sets: CATH S-Level, CATH H-Level, SCOP family level and SCOP superfamily level, and

we evaluate five family sets: CATH S-Level, CATH H-Level, SCOP family level, SCOP superfamily level and EVEREST. The results of the coverage and accuracy tests we have performed are shown in the figure. For all reference sets, the most covering evaluated sets is always EVEREST (on tie with SCOP superfamilies for covering CATH H-Level). Of the SCOP families, 32% are perfectly reconstructed by EVEREST and only 3% by CATH. Of CATH H-Level families, 31% are perfectly reconstructed by EVEREST and 27% by SCOP. Thus, EVEREST includes families of various granularities, covering all sorts of notions of the concept family. EVEREST performs almost as well as SCOP in terms of accuracy with respect to CATH S-Level, and almost as well as CATH in terms of accuracy with respect to SCOP families. It is less accurate with respect to CATH H-Level and SCOP superfamilies.

We conclude that EVEREST domain families provide a good reconstruction of more than one hierarchical level and that the inconsistency between SCOP and CATH is similar than between EVEREST and each of these classifications.

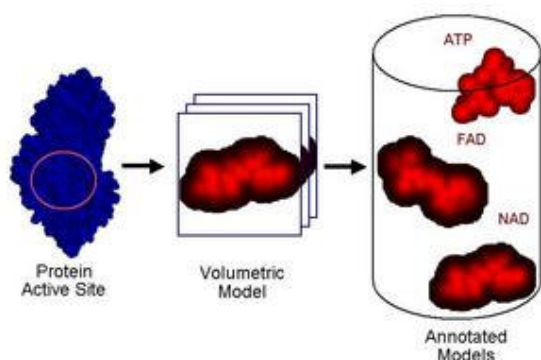
References:

- [1] H. M. Berman, J. Westbrook, et al. The Protein Data Bank. *Nucleic Acids Res*, 28:235-42, 2000.
- [2] J. M. Chandonia, N. S. Walker, et al. ASTRAL compendium enhancements. *Nucleic Acids Res*, 30(1):260-3, 2002.
- [3] C. Hadley and D. T. Jones. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, 7(9):1099-112, Sep 1999.
- [4] T. J. Hubbard, B. Ailey, et al. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res*, 27(1):254-6, 1999.
- [5] C. A. Orengo, F. M. Pearl, et al. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res*, 27(1):275-9, 1999.
- [6] E. Portugaly, A. Harel, et al. EVEREST: Automatic identification and classification of protein domains in all protein sequences. <http://www.everest.cs.huji.ac.il/>

**Figure Legend: Accuracy and coverage tests for four reference sets and five evaluated domain family sets.** Line colors code for the different evaluated sets. Upper panels report accuracy tests with respect to different reference sets. Lines describe distributions of scores of evaluated families, e.g. the height of the right end of the line is the proportion of evaluated families perfectly reconstructing a reference family. Lower panels report coverage tests with respect to different reference sets. Lines describe distribution across reference families of scores of best matching evaluated family, e.g. the height of the right end of the line is the proportion of reference families perfectly reconstructed. For coverage tests, in order to avoid trivial families, only reference families of size 5 or more, which appear in some protein in hetero-multi-domain context are considered.

**66 Thomas Funkhouser<sup>1</sup>, Roman Laskowski<sup>2</sup>, Janet Thornton<sup>3</sup>** Princeton U., <sup>1</sup>funk@cs.princeton.edu  
<sup>2</sup>European Bioinformatics Institute, roman@ebi.ac.uk <sup>3</sup>thornton@ebi.ac.uk

## Matching Volumetric Models of Protein Active Sites



**Short abstract:** This paper investigates the use of a knowledge-based algorithm to build volumetric models of active site cavities from protein structures, a fast rotational matching algorithm to detect similarities in the volumetric models, and a nearest neighbor classifier to predict the type of ligand bound based on similarities in volumetric models.

**Full abstract:** The goal of our project is to detect functional similarities between protein active sites. Given the 3D atomic coordinates for a protein, along with the location of an active site, we employ a knowledge-based algorithm based on X-Site [Laskowski96] to build a volumetric model of the chemical and geometric properties inside its cavity. We then use a fast rotational matching algorithm [Kovacs02] to find the correlation between pairs of volumetric models at the optimal rotational

alignment. Finally, we use a nearest neighbor classifier to transfer functional annotations between similar active sites.

To test the proposed methods, we performed a leave-one-out classification study aimed at predicting the type of ligand bound in active sites of proteins in the PDB. Our test set comprised 105 active sites of non-homologous proteins partitioned into 6 groups (ATP, FMN, BOG, NAD, FAD, and HEM). We constructed a volumetric model for each site, matched all pairs, and then measured how often the best match for each active site binds the same type of ligand. For comparison sake, we also computed the classification rate achieved with FASTA [Pearson90], (a sequence-alignment program), ICP [Besl92] (a method for aligning atoms in a 15 Å sphere surrounding the active site), CE [Shindyalov98] (a structure alignment program), SCOP [Murzin95] (a structural classification), and random (a random classifier).

We found that volumetric matching correctly predicts the bound ligand type for 61% of the active sites, as compared to 50% with SCOP, 47% with CE, 46% with FASTA, and 34% with a random classifier. These results suggest that volumetric matching of binding sites is useful for predicting gross categories of molecular function (e.g., bound ligand type). Further study is required to determine if they are also effective at distinguishing more specific molecular functions (e.g., EC number) and whether they can be used to discover the functions of novel proteins for which no function is currently known.



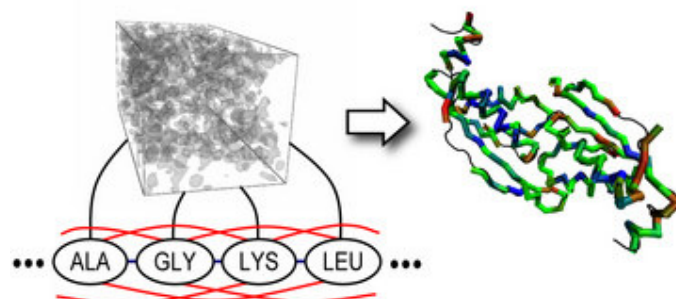
## REFERENCES

- [Kovacs02] J.A. Kovacs and W. Wriggers (2002). Fast rotational matching, *Acta Cryst.*, D58:1282-1286.  
 [Laskowski96] R.A. Laskowski et al. (1996). X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins, *J. Mol. Biol.*, 259:175-201.  
 [Murzin95] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247:536-540.  
 [Pearson90] W.R. Pearson (1990). Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol.*, 183:63-98.  
 [Shindyalov98] I.N. Shindyalov and P.E. Bourne (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng.*, 11:739-747.

**69 Frank P DiMaio<sup>1</sup>, Jude W Shavlik<sup>2</sup>, George N. Phillips** U. of Wisconsin - Madison, <sup>1</sup>dimaio@cs.wisc.edu  
<sup>2</sup>shavlik@cs.wisc.edu

### A Probabilistic Approach to Protein Backbone Tracing in Electron Density Maps

**Short abstract:** One particularly time-consuming step in x-ray crystallography is interpreting the electron density map. We



describe a probabilistic approach to automated backbone tracing in poor-quality density maps using a Markov random field. The model is flexible, considering an almost-infinite number of possible backbone conformations, while ignoring physically impossible ones.

**Full abstract:** There has been much recent interest in rapidly determining protein structures using X-ray crystallography. One particularly time-consuming step in protein crystallography is interpreting the electron density map. The electron density map –

analogous to a 3D image of the protein – is produced by the crystallographic process, and interpreting this map involves fitting a complete molecular model. In poor-quality electron density maps, the interpretation may require a significant amount of a crystallographer's time. Our work investigates automating the time-consuming initial backbone trace in poor-quality density maps. We describe ACMI (Automatic Crystallographic Map Interpreter), which uses a probabilistic model known as a pairwise Markov random field to represent the protein.

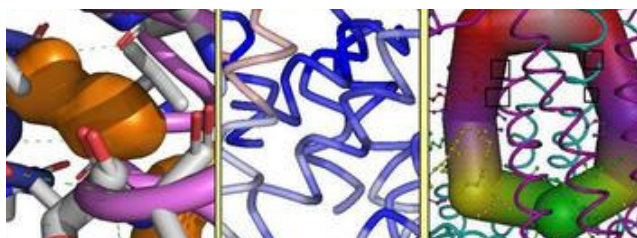
A pairwise Markov random field defines the joint probability of some set of random variables over an undirected graph. Specifically, the probability of a particular setting of the random variables in the graph is represented as the product of potential functions defined over edges, and over vertices in the graph. In modeling the protein backbone, random variables describe the three-dimensional positions and orientation of amino-acid residues in the electron density map. Vertices in the graph correspond to amino-acid residues. Potentials associated with each vertex are derived by matching a set of rigid templates – constructed from the associated residue – into the density map. Edges model spatial constraints within the protein. In ACMI, edge potentials fall into two types: adjacency constraints model interactions between adjacent residues, while occupancy constraints model interactions between residues distant on the protein chain (though not necessarily spatially distant in the folded structure).

Modeling the protein in this manner allows the model to be flexible, considering an almost infinite number of possible conformations, while rejecting any that are physically impossible. Using an efficient algorithm for approximate inference – loopy belief propagation (BP) – allows the most probable trace of the protein's backbone through the density map to be determined. Several enhancements, which allow BP to scale to proteins with thousands of residues, are also presented. This technique has yielded favorable results, providing accurate interpretation even in poor quality maps. We test ACMI on a set of ten protein density maps (at 2.5 to 4.0<sub>Å</sub> resolution), and compare our results to alternative approaches. At these resolutions, ACMI offers a more accurate backbone trace than current approaches.

**76 Ilan Samish<sup>1</sup>, Oksana Kerner<sup>2</sup>, David Kaftan, Eran Goldberg, Avigdor Scherz** <sup>1</sup>Present address: Dept. of Chemistry and Dept. of Biochemistry and Biophysics, U. of Pennsylvania, samish@sas.upenn.edu <sup>2</sup>Weizmann Institute of Science

### Datamining the Fourth Dimension from Crystal Structures: Structure-Function-Entropy Relationships in Membrane Proteins

**Short abstract:** Datamining crystal structures, we show how variation in local conformational flexibility, entwined with



structural motifs, controls the dynamic function of membrane proteins. Our experimentally validated analysis explains Biological dynamic mechanisms and provides new simple guidelines for the study of membrane proteins.

**Full abstract:** Crystal structures, the major source of protein structural information, are often regarded as

static 3D snapshots of dynamic macromolecules. Thus, study of protein dynamics is often confined to computationally heavy simulations. Instead, here we data-mine membrane-protein structures directly. We show how local regions of distinct flexibility are fine-tuned to control specific functions. Protein features considered during the analysis include intra-protein cavities, B-factors normalized in a unique manner, and intrinsic degrees of freedom. These are correlated with evolutionary conservation, packing motifs, mutagenesis simulations, local deformations, functional pathways, and experimental validation.

First, we show how local flexibility acclimatizes photosynthetic energy conversion across a wide range of temperatures, from frigid Alaska to scalding hot springs<sup>[1]</sup>. A group of hitherto unrecognized intra-protein cavities and adjacent packing motifs jointly impart localized flexibility at the center of the photosynthetic complex. The motifs are evolutionarily conserved at the sequence and structural levels with consistent changes solely in extremophiles. As validated experimentally in mesophiles and thermophiles, electron transfer rate accelerates with temperature (Arrhenius behavior) up till the physiological temperature, where it levels-off, thus avoiding photo-damage due to reaction over-acceleration. Moreover, *in silico* mutagenesis that abolished the cavity was applied *in vivo* and shown to eliminate the non-Arrhenius behavior.

Second, to generalize our findings, we assess whether local flexibility, associated with unique hydrogen bonding patterns, confers specific functions. As recently reviewed<sup>[2]</sup>, the frequent appearance of local distortions in transmembrane helices may enable precise positioning of functional groups or, in contrast, facilitate flexible functional movement. Data-mining a non-redundant dataset of crystal structures reveals that both roles of these local distortions are demonstrated. They are clearly divided according to the normalized temperature- (or B-) factor. In order to apply this measurement of crystal structure disorder to the local region of interest, normalization was confined to transmembrane backbone atoms. Further, in order to localize the helical deformations, intrahelical H-bonding parameters were utilized, rather than the conventional DSSP annotation. The approach provides guidelines for analysis of pivotal, evolutionarily conserved loci.

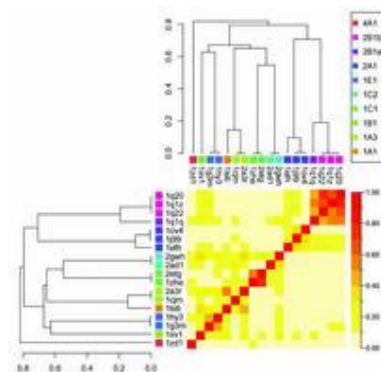
Last, we assess whether the protein matrix disorder level per se can modulate function? We show that variation in local flexibility along electron transfer pathways explains the dilemma between two opposing requirements: optimizing rigid geometry required for donor-to-acceptor wave-function overlap and conferring flexibility to quickly dissipate heat thus avoiding a back-reaction. In photosynthetic complexes the first, fast and low-energy-gap steps take part along a rigid microenvironment thus ensuring chemical capture of the photo-excited state. In contrast, the latter slow steps involving a high-energy-gap steps take part along a flexible microenvironment.

Cumulatively, data-mining the local variation in the intrinsic flexibility of crystal structures provides a tool to study dynamic mechanisms. The applied methods provide guidelines to analyze key functions, adding available, yet often neglected, data to our understanding of membrane proteins.

1. Kerner, O.\*, Samish, I.\*, Kaftan, D.\*, Holland, N., Sai, P. S. M., Kless, H. & Scherz, A. (2006). Protein Flexibility Acclimatizes Photosynthetic Energy Conversion to the Ambient Temperature. *Nature* (In Press). \*Equal contribution
2. Bowie, J. U. (2005). Solving the Membrane Folding Problem. *Nature* 438, 581-589.

**82 Rafael J Najmanovich<sup>1</sup>, Richard J Morris<sup>2</sup>, Janet M. Thornton** European Bioinformatics Institute, <sup>1</sup>rafael.najmanovich@ebi.ac.uk <sup>2</sup>Richard.Morris@bbsrc.ac.uk

## Single-Family Analysis of Binding Site Structural Similarities



**Short abstract:** We describe a method for the detection of pairwise local structural similarities between members of a given protein family using all non-hydrogen atoms in protein clefts allowing larger, over-predicted and apo-form binding sites to be analyzed. We uncover previously unknown similarities between members of the human cytosolic sulfotransferase family.

**Full abstract:** A renewed emphasis is being placed in recent years into family-wide approaches to structural genomics with the goal of reaching a better understanding of protein function in terms of molecular recognition and ligand binding.

While the overall fold as well as the general function is conserved within a given protein family, variations in specific positions within the fold or larger differences in loops can be responsible for subtle changes in the regulation or function of various members of the protein family and might be responsible particularly for differences in substrate specificity and selectivity. Furthermore, local structural

similarities in active/binding site regions, can provide a different perspective into the evolution of a protein than that obtained using the overall sequence or structure (Via, Ferre et al. 2000; Campbell, Gold et al. 2003). The study of such family-wide similarities is a crucial step in understanding protein function at the molecular level with practical implications in the field of drug design.

Methods for the detection of local (binding site) structural similarities have been developed in recent years primarily for the prediction of function from structure rather than the analysis of members of a given protein family. Such methods can be useful in providing hints about the function of a query protein particularly in cases where sequence-based methods as well as methods based on the overall fold of the protein are unable to give clues.

Methods for the prediction of function from structure via the detection of local structural similarities can be thought as

composed of three interconnected parts: structure representation, search method and scoring procedure. The various methods developed over the years can differ in any of these components. Furthermore, some methods perform the search of the query structure against pre-defined structural templates. These templates can be automatically generated or curated, in which case their biological relevance is known (Najmanovich, Torrance et al. 2005).

Given the sets of atoms defining the clefts under comparison, the question that needs to be answered is what is the largest subset of atoms in both clefts in direct correspondence with each other geometrically as well as chemically. This is a combinatorial optimization problem where, in principle, each possible set of atom correspondences might be a solution and the largest such set is the global solution. Graph theory offers a means to solve this problem via the detection of the maximal (largest) clique in an association graph. Further details on graph theory can be found elsewhere (Gross and Yellen 2004). In the present work, we use the standard algorithm of Bron and Kerbosh (Bron and Kerbosch 1973) for the detection of cliques.

Depending on the number of atoms being compared, the size of the association graph might make it practically unfeasible to detect the largest clique when considering all non-hydrogen binding site atoms. In order to overcome this difficulty we perform the graph matching in two stages.

In the first stage, an initial superimposition is performed via the detection of the largest clique in an association graph constructed using only C-alpha atoms of identical residues in the two clefts. A maximum distance difference of 2.0 Angstrom is used to create edges in the association graph, imposing an upper bound of the same magnitude in the coordinates root mean square distance (RMSD) of corresponding C-alpha atoms.

Once the largest C-alpha clique is obtained its transformation matrix and translation vector are used to superimpose all atoms in the two clefts using the least square method of Arun et al. (Arun, Huang et al. 1987) based on the singular value decomposition of the coordinates variance-covariance matrix.

In the second graph matching stage, all non-hydrogen atoms are used. Association graph nodes are created with the requirement that two atoms, one from each cleft, be of the same atom type as well as that their spatial distance be within 1.5 Angstrom. This spatial distance constraint is used to decrease the size of the association graph and is the reason why the initial superimposition is performed. In the present work, we use eight atoms type classes (Sobolev, Wade et al. 1996; Sobolev, Sorokine et al. 1999) comprising the following classes: Hydrophilic, Acceptor, Donor, Hydrophobic, Aromatic, Neutral, Neutral-donor and Neutral-acceptor. Similar to the first stage, a maximum distance difference in defining association graph edges is used. This second threshold is set to 1.5 Angstrom and again defines an upper bound of that magnitude in the RMSD between corresponding non-hydrogen cleft atoms.

The method for the detection of local structural similarities and the Tanimoto coefficients of sequence and structural similarity developed here are able in conjunction to detect binding site sequence and structural similarities between members of the human cytosolic sulfotransferase family. Binding site sequence and structural comparisons uncovered similarities between members of two subfamilies of the human cytosolic sulfotransferase family not seen through overall sequence comparisons.

Arun, K. S., T. S. Huang, et al. (1987). "Least-Squares Fitting Of 2 3-D Point Sets." *Ieee Transactions On Pattern Analysis And Machine Intelligence* 9(5): 699-700.

Bron, C. and J. Kerbosch (1973). "Algorithm 457: finding all cliques of an undirected graph." *Communications of the ACM* 16(9): 575-577.

Campbell, S. J., N. D. Gold, et al. (2003). "Ligand binding: functional site location, similarity and docking." *Current Opinion in Structural Biology* 13(3): 389-395.

Gross, J. L. and J. Yellen (2004). *Handbook of graph theory*. Florida, CRC Press.

Najmanovich, R. J., J. W. Torrance, et al. (2005). "Prediction of protein function from structure: insights from methods for the detection of local structural similarities." *Biotechniques* 38(6): 847, 849, 851.

Sobolev, V., A. Sorokine, et al. (1999). "Automated analysis of interatomic contacts in proteins." *Bioinformatics* 15(4): 327-332.

Sobolev, V., R. C. Wade, et al. (1996). "Molecular docking using surface complementarity." *Proteins-Structure Function and Genetics* 25(1): 120-129.

Via, A., F. Ferre, et al. (2000). "Protein surface similarities: a survey of methods to describe and compare protein surfaces." *Cellular and Molecular Life Sciences* 57(13-14): 1970-1977.

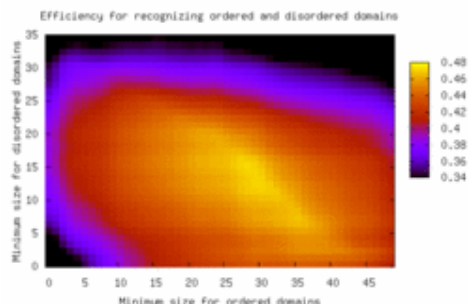
**84 Zsuzsanna Dosztanyi<sup>1</sup>, Márk Sándor<sup>2</sup>, István Simon** Institute of Enzymology, Budapest, Hungary,  
<sup>1</sup>zsuzsa@enzim.hu <sup>2</sup>sanmark@freemail.hu

## Prediction of Order and Disorder at the Domain Level

**Short abstract:** During the target selection process of structural genomics projects it is important to be able to discern ordered globular domains from disordered regions. A method based on the IUPred algorithm is suggested to locate ordered and disordered domains from amino acid sequences and was tested on specific datasets.

**Full abstract:** Intrinsically unstructured/disordered proteins (IUPs) lack a stable 3D structure, rather exist in highly flexible, unfolded structural state. Disordered regions often lead to difficulties in protein expression, purification and crystallization. Therefore, it is important to be able to discern ordered globular domains from disordered regions during the target selection process of structural genomics projects. Existing prediction methods assign order/disorder at the level of individual positions and their performance was tested using fully ordered or disordered regions. The task of predicting





order and disorder at the level of domains, however, requires different type of datasets, algorithms, and evaluation criteria. Only a few methods have been tuned for assign order and disorder at domain level [1,2,3], but no systematic study is available.

The main challenge here is to find the boundaries between disordered and ordered regions, without predicting many short segments. Compared to the observed domain sizes, predictions often contain only short continuous stretches of ordered or disordered positions. Noise in the prediction methods can be one source of such unavoidable fragmentation. However, short regions can also occur when the prediction methods realistically capture some local tendencies for order or disorder. In most cases, these regions should be overlooked aiming at prediction at the domain level. For

example, many X-ray structures contain regions with missing electron density, which do not prevent overall structure determination. Consequently, they should be merged into globular domains. The opposite case, short ordered regions within large disordered regions were also observed and were suggested to correspond to functional sites. Such regions are not capable of forming stable structures in themselves and should be considered as part of disordered regions. However, the exact requirements to form globular domains or disordered segments are not known.

The optimal rules for discerning domains and regions for both order and disorder were explored via various algorithms. The predictions were based on the IUPred [4] that estimates the energetic contribution of each residue from the amino acid composition of its neighborhood. It assumes that lack of 3D structure for IUPs is the result of their specific amino acid composition which does not allow sufficient favorable interactions to form. The strength of this approach is that its parameters are based on globular proteins only, without relying on datasets of unstructured proteins. Starting from IUPred predictions, various algorithms have been implemented in order to tune the performance for recognizing ordered domains and long disordered segments. Each algorithm can be described by a few basic rules and required only a few parameters. Some of these were based on the length of regions. Taking advantage of the physical model behind IUPred, the energy of regions can also be used to decide which regions should be treated self-contained.

In the first algorithm we followed the procedure applied in GlobPlot [2]. This eliminates short segments based on their length only. Neighboring globular domains were merged if the intervening disordered region was too short, and ordered regions below some minimal size were omitted. In another algorithm an iterative procedure was implemented. Regions were ranked according to their lengths, and short regions were reassigned to their neighborhood. Different size limits were used for ordered and disordered regions. A similar algorithm used the energy instead of the length of segments for ranking and merging. In the last set of algorithms we tried to find directly the starting and end position of segments with the most favorable energy. The search was repeated until non-overlapping segments could be found in the sequence.

The evaluation of these algorithms requires a specific dataset containing long regions of both order and disorder, mixed the same way as observed in proteins. The performance of the algorithms was tested on two datasets. Only a small dataset could be collected from experimentally verified disordered regions which were juxtaposed by domains with known structure. A much larger dataset was generated using the mostly ordered domains of the Pfam database [5]. Regions containing both a Pfam family and an unassigned putative disordered region were selected for the purpose of optimizing the parameters and testing the performance of the various algorithms. Unlike in domain prediction, consecutive globular domains were treated as a single ordered unit. Although this larger dataset is less reliable, the performances and the optimal parameters were similar to that of the small dataset.

[1] Oldfield CJ, Ulrich EL, Chen Y, Dunker AK, Markley JL (2005) *Proteins* 59:444-453

[2] Linding R, Russell RB, Neduva V, Gibson TJ (2004) *Nucleic Acids Res.* 31: 3701-3708.

[3] Dosztányi Zs, Csizmadik V, Tompa P, Simon I (2005) *Bioinformatics* 21:3433-3434

[4] Dosztányi Zs, Csizmadik V, Tompa P, Simon I (2005) *J. Mol. Biol.* 347:827-839.

[5] Bateman A, Coin L, Durbin R, Finn RD, et al. (2004) *Nucleic Acids Res.* 32:D138-D141.

## Laptop Session Abstracts

**11 Jorge H Fernandez<sup>1</sup>, Solange M Serrano<sup>2</sup>, Wilson Savino, Ana Tereza Vasconcelos**

<sup>1</sup>LabInfo, Laboratório Nacional de Computação Científica/MCT, Petrópolis, R.J, Brazil, jorgehf@lncc.br <sup>2</sup>Center of Applied Toxinology – Instituto Butantan, S.P., Brazil, serrano@butantan.gov.br

**Bothrostatin, a new RGD-type disintegrin from Bothrops jararaca venom: binding specificity in bothrostatin- $\alpha$ IIb $\beta$ 3/ $\alpha$ v $\beta$ 3 integrin complexes.**

**Short abstract:** Disintegrins are among the most potent antagonists of several integrin receptors, blocking specific integrin-ligand interactions in important biological processes. We use molecular dynamics to study the integrin/ligand interaction specificity in  $\alpha$ IIb $\beta$ 3/ $\alpha$ v $\beta$ 3-disintegrin complexes. We found that C-terminal tail of disintegrin structure is crucial for integrin recognition and binding affinity.

**Full abstract:** Disintegrins are among the most potent antagonists of several integrin receptors, blocking specific integrin-ligand interactions in important biological processes as thrombosis,



angiogenesis, cancer and inflammation. We described the structure, integrin binding specificity and dynamics of inhibition of Bothrostatin, a novel RGD-type disintegrin from Bothrops jararaca venom that showed an nM inhibitory activity on collagen-induced platelet-aggregation.

Comparative modeling approach was used to obtain a structural model of bothrostatin in complex with  $\alpha$ IIB $\beta$ 3 and  $\alpha$ v $\beta$ 3 integrins, and molecular dynamics of the complexes were compared as a useful tool for studies of integrin/ligand interaction specificity. In energy minimization, equilibration and production dynamics was used GROMACS and interactions in final structures were evaluated in STAR Sting package.

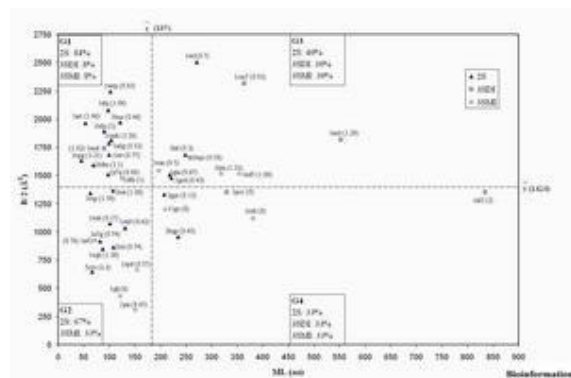
Our results suggest that bothrostatin adopts a globular, closed structure in physiological solution. The RGD motif is exposed to the solution by the loop formed by 45-59 residues and is very close to the C-terminal domain forming a finger-like structure. General recognition of the integrin-disintegrin complex is driven by electrostatic interactions, but in final ligand accommodation and binding affinity, the C-terminal tail of disintegrin structure enlarge the recognition surface contacting with  $\alpha$  integrin subunit. These “in silico” results are in concordance with the fact that integrins recognizing the same RGD sequence in their ligands ( $\alpha$ IIB $\beta$ 3,  $\alpha$ v $\beta$ 3 and  $\alpha$ 5 $\beta$ 1) show different binding affinity towards different RGD-disintegrins. These results will help us to understand the dynamics of the integrin-ligands binding and recognition, and to develop new, potent and integrin-specific inhibitors.

Supported by CNPq, MCT and FAPESP.

**12** Lie Li<sup>1</sup>, Meena Sakharkar<sup>2</sup>, Pandjassaram Kanguane<sup>3</sup>, Jacob Gan, Pandjassaram

Kanguane NTU, <sup>1</sup>PG03074376@ntu.edu.sg <sup>2</sup>mmeena@ntu.edu.sg <sup>3</sup>mpandjassaram@ntu.edu.sg

**Structural features differentiate the mechanisms between 2S (2 state) and 3S (3 state) folding of homodimers.**



**Short abstract:** The formation of homodimer complexes for interface stability, catalysis and regulation is intriguing. Here, we analyze 41 homodimer (25 `2S` and 16 `3S`) structures determined by X-ray crystallography to estimate structural differences between them. The findings are useful in the prediction of homodimer folding and binding mechanisms using structural data.

**Full abstract:** The formation of homodimer complexes for interface stability, catalysis and regulation is intriguing. The mechanisms of homodimer complexations are even more interesting. Some homodimers form without intermediates (two-state (2S)) and others through the formation of stable intermediates (three-state ((3S))). Here, we analyze 41 homodimer (25 `2S` and 16 `3S`) structures determined by X-

ray crystallography to estimate structural differences between them. The analysis suggests that a combination of structural properties such as monomer length, subunit interface area, ratio of interface to interior hydrophobicity can predominately distinguish 2S and 3S homodimers. These findings are useful in the prediction of homodimer folding and binding mechanisms using structural data.

Methods: Dataset creation: We created a dataset consisting of 41 homodimer complex structures (2S: 25; 3SDI: 5; and 3SMI: 10) from Protein Databank (PDB). [8] The unfolding pathways for these dimers observed using thermodynamic experiments were obtained from literature (Table 1). The selected homodimers are at least 40 residues per monomer. Analyses of 2S and 3S homodimers Interface area: The solvent accessible surface area (ASA) was computed using the program NACCESS. [9] The dimeric interface area (B) was calculated as  $\Delta$ ASA (change in ASA upon complex formation from monomer to dimer state). [10] We then calculated subunit interface area (B/2), due to the twofold symmetry of homodimer complexes.

Interior, interface and exterior residues: Homodimer residues were classified into three categories (interior, interface and exterior) based on relative ASA. The percentage relative ASA was obtained by dividing the accessible surface area by the total surface area of a side-chain in an extended conformation in the tripeptide GXG. Exterior residues were defined as having a relative ASA > 5%, interior residues were defined as having a relative ASA < 5% and interface residues were defined satisfying the conditions  $\Delta$ ASA > 12\_ & relative ASA < 5%. The 5% cut-off was optimized elsewhere by Miller et

al. [11]

Fraction of interface to interior Hydrophobicity (Fhp): Fhp (Fraction of interface to interior hydrophobicity) was defined by the equation  $(\text{Hinf-Hext})/(\text{Hint-Hext})$ , where Hint is interior hydrophobicity, Hinf is interface hydrophobicity and Hext is exterior hydrophobicity. The individual hydrophobicity values were calculated using the equation  $\sum n_i h_i / \sum n_i$ , where  $n_i$  is the number of residue type  $i$  and  $h_i$  is hydrophobicity value (based on SES (solvent excluded surface) & SAS (solvent accessible surface)) of type  $i$ , as described elsewhere. [12]

Small and large homodimers: By definition, small homodimers were defined as those with ML (monomer length) less than the dataset mean length (185 residues). By definition, large homodimers were defined as those with ML larger than the dataset mean length (185 residues).

Homodimers with small and large B/2: By definition, homodimers with small B/2 were defined as those whose B/2 is less than the dataset mean B/2 (14242<sub>Å</sub><sup>2</sup>). By definition, homodimers with large B/2 were defined as those whose B/2 is larger than the dataset mean B/2 (14242<sub>Å</sub><sup>2</sup>).

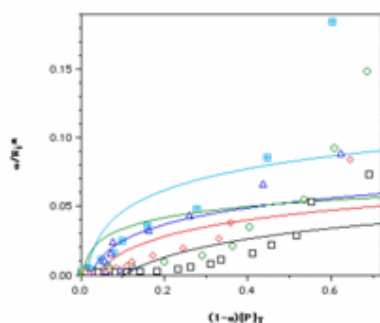
Conclusions: The mechanisms of homodimer complexations have implications in drug discovery. However, elucidation of homodimer mechanism using unfolding experiments is difficult. Prediction of homodimer folding and binding using structural data has application in target validation. Here, we show that small proteins with large interface area and high Fhp form 2S. We also show that large proteins with small interface area and low Fhp form 3S. Therefore, it is feasible to differentiate 2S and 3S homodimers using structural data.

Figure Legend: Correlation between monomer length (ML) and subunit interface area (B/2) for three groups of homodimers. 2S: two-state; 3SDI: three-state with dimeric intermediate; 3SMI: three-state with monomeric intermediate. The two dash lines through 185 aa and 14242<sub>Å</sub><sup>2</sup> represent mean monomer length and mean B/2 for all homodimers, respectively. They classify the dimers into four regions (G1, G2, G3 and G4). The distributions of 2S, 3SDI and 3SMI dimers are given for each region. The value within parentheses is hydrophobicity factor (Fhp), calculated by the equation  $(\text{Hinf} - \text{Hsurf})/(\text{Hint} - \text{Hsurf})$ , where Hinf is interface hydrophobicity, Hint is interior hydrophobicity and Hsurf is surface hydrophobicity.

### 13 Francisco Torrens<sup>1</sup>, Gloria Castellano<sup>2</sup>

<sup>1</sup>Universitat de Valencia-Institut Universitari de Ciència Molecular, Francisco.Torrens@uv.es <sup>2</sup>Universidad Politécnica de Valencia-Universidad Católica de Valencia, gcaste@trovador.zzn.com

### Binding of Water-Soluble, Globular Proteins to Anionic Model Membranes



**Short abstract:** The electrostatics role in the adsorption of proteins to negatively-charged/neutral vesicles is studied. The interaction is monitored at low ionic strength for lysozyme/myoglobin/albumin as a function of pH. Adsorption is investigated via. fluorescence emission spectrum changes. Adsorption is found at pHs where the protein charge is at least 3 e.u.

**Full abstract:** The role of electrostatics in the adsorption process of proteins to performed negatively-charged (phosphatidylcholine/phosphatidylglycerol, PC/PG) and neutral (PC) small unilamellar vesicles (SUV) is studied. The interaction is monitored at low ionic strength for a set of model proteins as a function of pH. The adsorption behaviour of lysozyme (pI = 10.7),

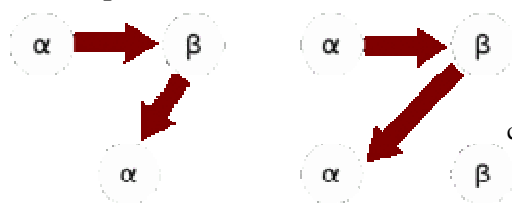
myoglobin (pI = 7.2) and bovine serum albumin (BSA, pI = 4.7) with preformed SUVs is investigated, along with changes in the fluorescence emission spectrum of charged proteins via the adsorption on SUVs. Significant adsorption of the proteins to negatively-charged SUVs is only found at pH values where the number of positive charge moieties exceeds the number of negative charge moieties on the protein by at least 3 e.u. Negligible adsorption to SUVs composed of zwitterionic lipids is observed in the tested pH range (4-9). The fluorescence emission of positively-charged proteins increases after adsorption on negatively-charged SUVs. With increasing protein to phospholipid ratio, the increase in the fluorescence emission levels off and reaches a plateau; protein adsorption profiles show a similar shape. Analysis of the data demonstrates that neutralization of the SUV charge due to the adsorption of the positively-charged proteins is the controlling factor in their adsorption. The reached plateau level depends on the type of protein and the pH of the incubation medium. This pH dependency can be ascribed to the mean positive charge of the protein. The effective charge of lysozyme, myoglobin and BSA (defined as the number of phosphatidylglycerol groups neutralized by one adsorbed protein molecule) is calculated from the charge differences between empty SUVs and protein-coated SUVs using the Gouy-Chapman theory. For lysozyme and myoglobin, an excellent correlation is found between the effective charge and the mean protein charge. The knowledge of molecular details of the interaction between lysozyme and membrane phospholipids is of interest for two reasons. (1) Lysozyme provides a convenient and suitable model for studying lipid-protein interactions involving initial electrostatic binding followed by putative penetration into the lipid bilayer. (2) The lysozyme-lipid interaction may provide important clues to understand the bactericidal action. Lysozyme, which is positively charged at pH 7.0, produces aggregation and fusion of negatively charged SUVs. These results are consistent with the ability of electrostatic charges, in protein binding to lipid SUVs, to produce the phenomena studied. The protein acts via a conformational change produced

upon binding to the lipid SUVs involving an increase in loops and unordered structure at the expenses of BETA-sheet conformation and BETA-turns. The forces involved in the interaction of lysozyme with lipid SUVs are not only electrostatic. In addition to electrostatic forces, membrane destabilization also plays a role in the aggregation and fusion of lipid SUVs induced by proteins. Lysozyme binding was demonstrated to alter the structure of the lipid phase of negatively charged membranes. Some results indicate the existence of hydrophobic interactions of the protein with the lipid bilayer. Strong experimental support for a hydrophobic component in the interaction is provided by the finding that not all bound lysozyme can completely dissociate from such membranes, either by increasing the ionic strength of the buffer solution, or by dilution of the bulk protein. Consistent with the electrostatic nature of the interaction, this effect occurs despite the fact that the initial binding itself is rather sensitive to both the ionic strength of the solution and the amount of the charge in the lipid membrane. The fact that electrically neutral lipid SUVs are able to bind the present protein evidently indicates an affinity based on hydrophobic interactions. Electrostatic interactions between the positive charges of lysozyme and negatively charged SUVs do also occur. Provisional conclusions follow. (1) The major role of electrostatics in the adsorption process of three relatively hydrophilic globular proteins and the physicochemical stability of SUVs adsorbed with proteins have been demonstrated. Adsorption to negatively-charged SUVs occurred only when the mean charge of the proteins was positive. Upon protein adsorption, the charge on the SUVs was reduced. Consequently, the driving force for further protein adsorption decreased. For lysozyme and myoglobin, the extent of charge neutralization as determined by spectrofluorimetry could be quantitatively ascribed to the mean charge of the proteins. (2) Lysozyme association with PC/CL membranes resulted in gradual decrease of fluorescence with protein concentration. Ionic-strength dependence of this effect suggested substantial contribution of electrostatic interactions to detected bilayer modifications. The initial binding step of lysozyme to the surface of the lipid membrane is governed by electrostatics. Subsequent to the initial stage of binding, changes in the structure of the protein and lipid components of the membrane occur, leading to a new and more complex situation involving penetration of lysozyme into the lipid phase.

**14 Karsten M Borgwardt<sup>1</sup>, SVN Vishwanathan<sup>2</sup>, Nicol N. Schraudolph, Hans-Peter Kriegel**

<sup>1</sup>Ludwig-Maximilians-U. Munich, kb@dbs.ifi.lmu.de <sup>2</sup>NICTA Australia, vishy@mail.rsise.anu.edu.au

### Fast Graph Kernels for Proteomics



**Short abstract:** Graph kernels are the principled approach to mining graph structured data, e.g. protein molecular structure descriptions. However, known graph kernels are too slow for large-scale applications on molecular graphs. We present fast graph kernels that can be used for large-scale protein function prediction or protein structure comparison.

**Full abstract:** Reaching from molecules to protein-protein-interaction networks, graph

structured data are very common in molecular biology. A general problem in bioinformatics is to measure similarity or to perform data mining on these graphs. Graph kernels are an attractive principled method for mining graph data, which can be used for protein function prediction, protein structure comparison or protein interaction-partner comparison. In practice, however, the application of graph kernels is hindered by the fact that their runtime complexity is too high for most large-scale applications and large graphs in bioinformatics. In ongoing research, we have already achieved a speed-up of the most prominent graph kernel by a factor of 1,000, making graph kernels fast enough for large applications in bioinformatics. In our talk, we would like to present how graph kernels work, how we managed to speed them up enormously and which problems in bioinformatics could be tackled by fast graph kernels. We are studying these problems in our current research.

Graph kernels are an attractive method for comparing graph data. While trivial approaches compare edges and nodes of two graphs only, and neglect valuable information stored in the graphs, other methods try to check if two graph contain common subgraphs - a very costly undertaking, as the runtime increases exponentially with the size of the graphs. Graph kernels offer a promising middle ground. They measure graph similarity by counting common substructures of graphs that are less costly to compute. These graph kernels have been successfully applied on several classification problems, e.g. on the MUTAG dataset (Gartner et al., COLT 2003).

The problem for the practitioner lies in the fact that even graph kernels tend to have a very high runtime. In ongoing research, we have studied the question how to speed up the computation of the random walk graph kernel by Gartner et al. (COLT 2003). Without losing accuracy of the solution, we have defined a novel approach for this graph kernel computation that is 1,000 faster than the direct approach. This progress is made possible by using Sylvester equations, conjugate gradient and fast matrix-vector multiplications.

With these fast graph kernels, it is now possible to measure similarity between large datasets of large graphs; these can be e.g. databases of 3D protein structures, or databases of protein interaction networks.

Graph kernels on proteins can be combined with Support Vector Machine Classifiers to create a successful protein function prediction system. In Borgwardt et al. (Bioinformatics, 2005), graph models of proteins that comprise structural and sequence information are compared via graph kernels to predict their enzymatic character.

Another interesting application that might gain importance over coming years is whole-network comparison of large protein-protein interaction (PPI) networks. This could be performed on PPI networks of different species to measure

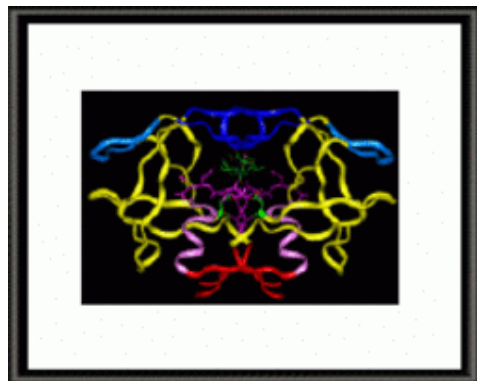


evolutionary relatedness, or between PPI of individuals - once available - to measure similarity of PPI in patients with different disease status.

Conclusions: In our presentation, we plan to demonstrate the potential of fast graph kernels in bioinformatics, proteomics, and 3D protein structure comparison in particular, and to encourage interest in and discussion about their emerging applications in bioinformatics.

**16 Talapady N Bhat<sup>1</sup>, Anh D Thi Nguyen, Alexander Wlodawer, Mohamed Nasr** NIST,  
<sup>1</sup>bhat@nist.gov

### Semantic Web and Chemical Ontology for Fragment-Based Drug Design and AIDS



**Short abstract:** Effective use of available structural information is crucial for fragment based design or fragment mix and match approach to succeed in drug discovery. Web based methods and chemical ontology for annotation and presentation of structural data for this purpose with specific emphasis to AIDS and Semantic Web will be presented.

**Full abstract:** The popular paradigm in fragment-based drug discovery seeks to collect, compare, and test many chemically similar compounds. Success of such a method critically depend on three factors; 1) how complete and reliable the chemical ontology is, 2) how well the compounds in the database have been annotated, classified, and indexed to produce meaningful and manageable hits, 3) what is the

relevance of steps (1) and (2) with respect to enzyme-drug interactions in the real world. 3-D structures narrate the secret stories of enzyme-drug interactions and they unravel the nature's secrets of drug resistance. Scientists use these stories to discover new drugs and to develop strategies to combat drug resistance. Solving 3-D structures is an expensive process; for this reason, only a small fraction of the over 1000 2-D structures have been solved in 3-D. A successful use of this vast amount of information on 2-D structures for drug-development requires efficient techniques and standards to cross-index 2-D and 3-D structures and their fragments. We have developed a novel fragment-based technique called Chem-BLAST (Chemical Block Layered Alignment of Substructure Technique) to dynamically establish structural relationships among compounds using their fragments. Using this method, for the first time we have enabled the users of HIVSDB to criss-cross between 2-D and 3-D data for drug design purposes. This novel fragment-based technology readily circumvents the difficulty faced by medicinal chemists in taking advantage of 2-D structural data to understand their mode interaction with the HIV enzymes. HIVSDB (<http://xpdb.nist.gov/hivpdb/hivpdb.html>) addresses all these issues through the development and deployment of state of the art fragment-based techniques and tools for the 21st century. HIVSDB has over thousand 2-D structures of potent HIV protease inhibitors collected mostly from published literature and it has about 50% more 3-D structures of HIV protease inhibitor complexes than found in the PDB. HIVSDB uses a unique indexing and annotating technique for inhibitors that allows the organization of the information into a chemical data relationship [http://xpdb.nist.gov/hivpdb/advanced\\_query\\_files/slide0002.htm](http://xpdb.nist.gov/hivpdb/advanced_query_files/slide0002.htm) or [http://xpdb.nist.gov/hiv2\\_d/advanced\\_query\\_files/slide0002.htm](http://xpdb.nist.gov/hiv2_d/advanced_query_files/slide0002.htm) ( See fig below) that may be readily translated into Semantic Web concepts either through 'text strings' (pyridine-2-sulfonamide-or-2-pyridinesulfonamide is-type pyridine)

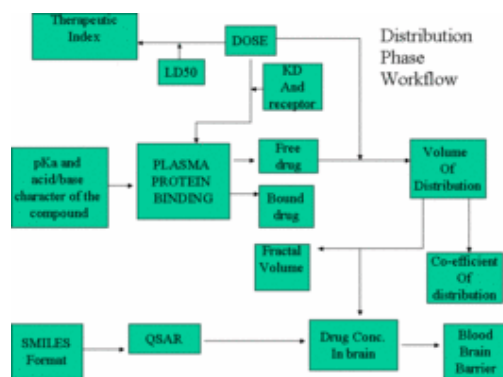
or 'visual tools' using the molecular images as shown in our web site

**16 Garima Goel<sup>1</sup>, zaharul ahsan<sup>2</sup>, Srinivasa Reddy Gurram** HCL Technologies, <sup>1</sup>garima.goel@hcl.in  
<sup>2</sup>mahsan@hcl.in

### Development of an effective model for in-silico prediction of distribution parameters of ADMET for viable drug molecules.

**Short abstract:** The model to define efficient distribution of the drug to its effector site is based on SAR and Statistical approaches to calculate Vd, CD, BBB and plasma protein binding.

**Full abstract:** With the increased cost to bring a drug molecule to market, the pharmaceutical industry has to improve the success rate of drug discovery and the various stages of drug development. The discovery and optimization of new drug candidates is becoming increasingly reliant upon the combination of experimental and computational approaches related to drug Absorption, Distribution, Metabolism, Elimination and Toxicity (ADMET). The in-silico model aims to predict the

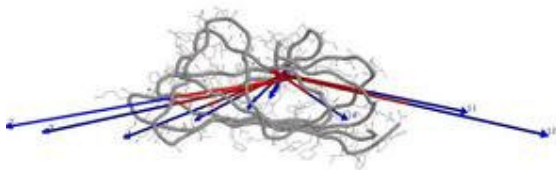


pharmacokinetic parameters of distribution phase (ADMET) such as volume of distribution (Vd), plasma protein binding, blood-brain barrier(BBB), co-efficient of distribution (CD), amount of drug (A), fractal volume of drug distribution (Vf) and therapeutic index which would help us to quickly evaluate the viable drug molecules .

Indigenous tool is based upon structure activity relationships (SAR) and other statistical based approaches which can be used for evaluating the possibility of a lead molecule to get effectively well – distributed throughout the body, targeting the receptor sites, efficiently binding to the tissues and organs, ability to cross the BBB and its safety index. The model can be used in early stages of drug discovery to assess the human distribution for potent novel molecules to achieve the goal of lead optimization with proper testing done for a set of anti-depressant drugs.

**18 Alexander A Kantardjiev<sup>1</sup>, Boris P Atanasov<sup>2</sup>** Bulgarian Academy of Sciences, Institute of Organic Chemistry,  
<sup>1</sup>alexkant@yahoo.com <sup>2</sup>boris@orgchm.bas.bg

### Electrostatics in Protein Docking: pH-Dependent self-consistent orientation guide



**Short abstract:** The protein electrostatic interactions are the only source for natural long-range interaction and are determinants of molecular recognition. Modern methodologies don't treat properly pH-dependence and self-consistency of charge interactions. Our analysis of solved protein complexes proves self-consistence between charge systems. It is how Nature suggests docking methodology.

**Full abstract:** It is well known that protein

electrostatic interactions (EI) play an essential role in protein structure-function relationships [1]. It is the only source for natural long-range interaction and is major determinant first step of an molecular recognition - "EI-enhanced diffusion" [2], protein-membrane interaction [3], immunoglobulin-target binding etc. Besides these fundamental issues, proper accounting for EI is critical for computational modeling of protein complexes and thus finds its application in protein and drug design. Nevertheless all docking methodologies are founded on shape complementarities and append simple Coulomb electrostatics, which does not treat properly pH-dependence and self-consistency of charge interactions. Electrostatic analysis of solved oligomeric protein complexes proves pronounced self-consistence between charge systems of subunits. As a result pKa-s values, degree of protonation, local electrostatic potentials (EP) of a subunit are influenced by neighbor subunits. So it is how Nature suggests docking methodology.

Following this conception we have designed and implemented novel protein docking strategy. We focus on two issues - pH-dependence of EI and self-consistency in charge interactions between subunits within the complex. The main aim in protein electrostatics [1] is to find 3D-EP distribution and its dependence on pH, ionic strength (Is), bound ions etc. We employ a fast iterative procedure for calculation of pH-dependent EI in proteins [4]. It accounts for all major terms in an EI procedure – model proton affinities, Born energy term, partial charges and pH-dependent proton charges. An important global characteristic of pH-dependent 3D-EP distribution is the electric/dipole moment (EP fitted to charges), which is basic determinant of long distance orientation in docking. In the assembling the subunits mutually influence their EI-s. This leads to self-consistent reorientation of electric moments along assembly pathway.

Following Nature wisdom we designed and implemented a novel docking procedure. Its basic characteristic is the leading role of electrostatics and self-consistency in complex formation. We follow two approaches: 1. Conventional search scheme (combination of Katzir-Katchalski and our pH-dependent EI method) and 2. Dipole guided assembly pathway (DGAP). In the first case the search procedure is formally similar to known docking schemes but the account for EI is different. As mentioned above we account for self-consistent EI. An additional advantage is that we can analyze docking at different pH and Is values. In the second case the DGAP docking procedure follows the oncoming molecules and their mutual orientation. The algorithm leads the molecules from their initial

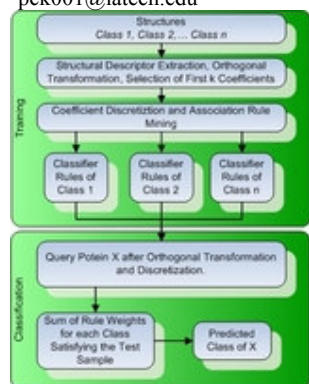
orientation through the assembling pathway by accounting for electric moments changes of both partners. Upon contacting the program tries to find the balance of electrostatic and contact interactions.

This approach differs most of the present-day docking schemes by allowing to evaluate pH-dependent effects: extreme pH values at which docking is not possible; pH intervals at which different complexes may form (by orientation); to follow the influence of ionic strength and many others. It can also give insight in the mechanism of assembly pathways. The approach can have diverse application beyond standard docking – from binding of ions/small ligands to adsorption on charged surface and electrodes.

References:

1. Honig, B. and Nicholls, A. (1995) Classical electrostatics in biology and chemistry. *Science*, 268, 1144-1149.
2. Getzoff, E.D., Tainer, J.A., Weinr, J.K, Kollman, P.A., Richardson, J.S. and Richardson, D.C. (1983) Electrostatic Recognition between Superoxide and Copper,Zinc Superoxide Dismutase Nature (London) , 306, 287-283.
3. Felder, C.E., Botti, S.A., Lifson, S., Silman, I. and Sussman, J.L. (1997). External and internal electrostatic potentials of cholinesterase models. *J. Molec. Graphics & Modelling* 15, 318-327
4. Kantardjiev A. and Atanasov B (2006) PHEPS: web based tool for pH-dependent Electrostatics of Proteins Nucleic Acid Research (accepted)

**20 Sumeet Dua<sup>1</sup>, Praveen C Kidambi<sup>2</sup>** College of Engineering and Science, Louisiana Tech U., <sup>1</sup>sdua@coes.latech.edu  
<sup>2</sup>pck001@latech.edu



## Protein Structural Classification using Associative Discrimination of Orthonormal coefficients

**Short abstract:** In this research, we present a novel automated computational framework for three dimensional (3D) structure-based classification of proteins using orthonormal transformation of the protein geometric shape descriptors, subsequently employing an association rule discovery-based supervised clustering approach to classify proteins.

**Full abstract:** Background: One of the daunting challenges facing Biology, and consequently multidisciplinary research in Computer Science, is to assign biochemical and cellular functions to the thousands of hitherto uncharacterized gene products discovered by several international gene-sequencing projects. These research endeavors produce protein amino-acid sequences of unknown function at an unprecedented rate. Proteins serve as some of the major structural elements of living systems within which their interactions determine most of the molecular and cellular operations. The latest release of the Protein

Data Bank (PDB) contains more than 36,200 proteins (April 2005) and records their three dimensional (3D) structures determined experimentally using primarily X-ray and NMR techniques. Although the number of such protein structures is increasing at a staggering rate, the number currently available remains trifling relative to the (over) 3.2 million known protein sequences (PIR-NREF, December 2005). Yet, due to the efforts of several international genomics initiatives, the number and availability of various protein structures have been increasing at a nearly exponential rate, leading to the demand for the development of fast, automatic, robust techniques of protein structure comparison, classification, modeling, and function prediction.

Motivation: Protein structure classification and comparison has become a central area in the field of bioinformatics. Rapid increase in size of protein databases has prompted the development of fast, automated methods to classify unknown protein structures. The resulting protein structural databases commonly suffer from the ‘curse of dimensionality,’ necessitating the development of methods of dimensionality reduction prior to classification. Moreover, design and development of efficient manual or semi-automated classification techniques has not kept pace with the growth in such databases. In this research, we present a novel automated computational framework for 3D structure-based classification of proteins using orthonormal transformation of the protein geometric shape descriptors, subsequently employing an association rule discovery-based supervised clustering approach to classify proteins.

Focus: Our research aims to address the following two questions: (a) Can a pair of novel geometric parameters (also called structural descriptors), obtained by observing the synthesis of protein structural subunits, be employed for effective dimensionality reduction of protein structures? (b) Can the inherent associations between such parameters be discovered and exploited to achieve accurate and fast protein structural classification? Our research involves the development of a novel computational framework for classifying protein structures using discrete cosine transformation (DCT) and class-association rules. The framework is presented in the attached figure.

Methodology: The proposed framework is divided into two basic phases: that of training and of classification. Steps involved in these phases were as follows:

1. The 3D structure of proteins is represented in terms of distributions of two previously proposed structural descriptors; namely, the  $NC_\alpha NC_\alpha$  dihedral angle and bond distance between the midpoints of the  $NCNC$  bond length between pairs of amino acids in a protein.

2. Orthonormal transformation using DCT is then applied to each of the derived distributions to represent them in the frequency domain. Since most of the distribution information is stored in the first few coefficients, it is only necessary to retain the first few  $k$  coefficients. This process assists dimensional reduction significantly due to the reduced number of  $k$  coefficients representing a protein relative to its original distribution length.

3. The transformed distributions of all proteins with known classes are merged into a dataset, which is given as input to the association rule-mining algorithm to generate *class-association rules* satisfying minimum *support* and *confidence*.

4. Final classifier rules for each class are selected based on their Laplace Expected Accuracy Error Estimate values. The test (or query) protein is then compared against the selected rules. Weights are assigned to each rule based on its size, confidence, and the coefficients involved in the rule (initial coefficients are given higher priority). Weights of rules satisfying the query protein are summed for each class and the class having the maximum weight is assigned to the query protein.

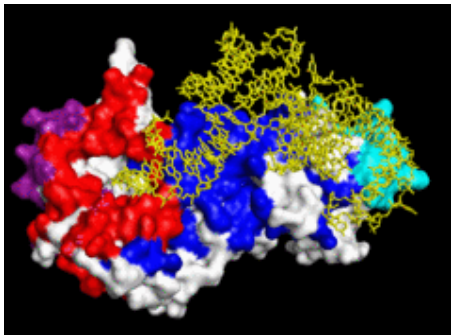
**Results:** We have demonstrated our results on two different datasets. The first balanced dataset contained 400 proteins from 10 different families. The 3D protein structure information was extracted from the PDB files, referred family-wise from the SCOP database. We experimented with 1-dimensional (1D) and 2-dimensional (2D) versions of DCT and established a higher classification accuracy (over 85%) by employing 2D DCT. In our second experiment, we implemented our framework on a dataset of 600 proteins from 15 folds. On this dataset, our method demonstrated an overall accuracy of over 83%. Thus, the proposed novel computational framework demonstrates the applicability of association rule discovery-based classification of structural descriptors for protein fold classification with improved sensitivity.

**Conclusions:** In this research, we have presented a novel computational framework for structural classification of proteins. Our method employs association rule-based classification which has proven to be an effective approach for high quality classification using protein structures. We have also used an efficient orthonormal transformation technique to represent the protein structure distribution in the frequency domain, thus greatly reducing the dimensionality of the original distribution. DCT has been extensively used in image and signal compression for its strong energy compaction properties. Using it, we are able to generate unique profiles for each protein family in the form of class-association rules. These rules use only a few DCT coefficients to represent the original protein distribution in the frequency domain. We have also employed two structural descriptors that effectively conserve the conformation of the entire protein structure and that possess unique distributions for proteins with varying sequence identities from different families.

Our future work involves experimenting with other forms of torsion angles and bond distances. We can also tune and modify our rule evaluation and selection routines to improve the performance of the classifier. A better weight assignment technique for a rule could also be employed that considers other factors and characteristics affecting rule selection. Overall, our classification scheme has the potential to develop into a very useful and accurate protein structural classification tool when employed on very large protein databases.

**21 Shula Shazman<sup>1</sup>, Yael Mandel-Gutfreund<sup>2</sup>**<sup>1</sup>Technion-Israel Institute of Technology, Faculty of Biology, shulas@tx.technion.ac.il <sup>2</sup>yaelmg@tx.technion.ac.il

## A Machine Learning Approach for Predicting RNA-binding Proteins from their Three Dimensional Structure



**Short abstract:** We apply a machine learning approach for predicting RNA-binding proteins from 3D structures. The method is based on characterizing unique properties of electrostatic patches on the protein surface. It does not rely on sequence or structural homology and could be applied to novel proteins with unique folds and/or binding motifs.

**Full abstract:** RNA has gained much interest in the recent years, mainly as a result of the large genome sequencing and the discovery of many non-coding functional RNA. Mostly RNA is not found free in the cell and in many cases it functions in complex with proteins. Specific recognition of RNA by proteins is mediated by different protein domains and motifs. Following the structure genomic initiate we expect that many proteins with potential RNA-binding function will be solved. In this study we apply a machine learning approach (Support Vector Machine) to predict RNA binding proteins from

their three dimensional structure and detect their binding interface. Previously we have developed an automated method for predicting protein function from structure and applied it successfully to predict DNA binding proteins (1). The method is based on characterizing the unique structural and sequence properties of large positively charged electrostatic patches on the protein surface. Here we apply a similar approach, concentrating on specific features which are characteristic to RNA binding proteins.

Using the vector of features extracted from the electrostatic patches on RNA and non-NA binding proteins we trained an SVM to recognize specifically RNA-binding proteins. Applying a hold out (jackknife) approach we showed that we can successfully distinguish the RNA-binding proteins from the rest of the proteins with a ROC value of 0.9. Interestingly,



among the false negative results there were 3 tRNA binding proteins, including tRNA-synthetase. Generally we found that the discriminating value generated by the SVM for each protein in the RNA-binding set correlated with the overlap between the largest positive electrostatic patch (calculated by our method) and the observed RNA-protein interface.

Overall we show that our method is successful in predicting RNA-binding proteins which are generally very diverse in terms of their structure, function and RNA recognition motifs. To the best of our knowledge the performance of the method is better than all other published methods for predicting nucleic-acid binding proteins. In addition, the method is much more efficient than our previous method for detecting DNA-binding protein as it does not rely on sequence conservation, which required time consuming calculations. Moreover, as the method does not require that the proteins have homologous sequence or structure in the databases, we suggest it could be applied to predict other RNA-binding proteins with novel folds and/or unique binding motifs.

1. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. Annotating nucleic acid-binding function based on protein structure. *J Mol Biol* 326, 1065-79 (2003).

**23 Oliver Sander<sup>1</sup>, Tobias Sing<sup>2</sup>, Francisco S. Domingues, Ingolf Sommer, Andrew Low, Peter K. Cheung, P. Richard Harrigan, Thomas Lengauer**<sup>1</sup>Max-Planck-Institute for Informatics, Saarbruecken,

Germany, osander@mpi-sb.mpg.de <sup>2</sup>tsing@mpi-sb.mpg.de

### Structural Descriptors Improve the Prediction of HIV Coreceptor Usage

**Short abstract:** Reliable methods for inferring HIV coreceptor usage are required for monitoring during antiviral therapy. Our novel structural descriptor considerably improves the predictive performance of purely sequence-based methods and is a first step to gain insights into structural determinants of coreceptor usage.

**Full abstract:** HIV cell entry requires a chemokine coreceptor in addition to the CD4 cell surface receptor. There are two types of coreceptors, CCR5 and CXCR4. Whereas CCR5-using viral variants dominate after infection and during early stages of the disease, in around 50% of the patients CXCR4-using variants appear in later

stages of the disease[6]. Although the causal relationship between this coreceptor switch from CCR5 to CXCR4 and the progression of disease is not clear yet, it is considered to be an important determinant of pathogenesis.

HIV coreceptors recently received increased interest as antiviral drug targets, with CCR5 antagonists being currently tested in clinical studies. Treatment with CCR5 antagonists requires continuous monitoring of coreceptor usage. Although phenotypic assays for determining coreceptor usage are commercially available, they are expensive and slow. To make monitoring coreceptor usage an element of routine clinical use, methods for inferring phenotype by assessing and interpreting genotypic information are desired. Although other parts of the viral envelope protein gp120 might subtly affect coreceptor usage, it is commonly agreed that the third variable loop region V3 is the major determinant for specificity and is therefore used in predictive methods.

Currently a simple charge rule based on amino acid charge at positions 11 and 25 in the V3 loop region[4] is routinely applied to infer coreceptor type. As the reliability of this sequence motif based prediction is limited[7], statistical learning methods based on support vector machines[5], neural networks[7], or position-specific scoring matrices[3] have been developed.

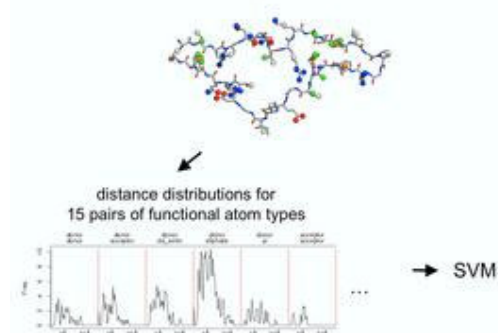
Whereas these current methods predict coreceptor usage based only on sequence information of the V3 loop, we use a recently published crystal structure of gp120 including the V3 loop[2] to exploit structural clues in prediction. By using this structure (PDB 2b4c, based on a CCR5-using JR-FL variant) as a rigid template of the V3 backbone conformation we compute structural descriptors and predict

the type of coreceptor used based on those. This approach improves the predictive performance of purely sequence-based methods and is a first step to gain insights into structural determinants of coreceptor usage.

**Methods & Materials:** Based on the loop backbone of PDB 2b4c we modeled the side-chain positions for each sequence variant using SCWRL[1]. The side chains are represented by functional atoms, labelled as hydrogen-bond donor, acceptor, ambivalent donor/acceptor, aliphatic, or aromatic ring according to [8]. For the subsequent statistical analysis the spatial arrangement is encoded by 15 distance distributions, one for each pair of functional atom types. Specifically, the pairwise distances between all types of functional atoms are modeled by kernel density estimates. The density estimates are uniformly sampled, resulting in a 15(distance distributions for atom type combinations) times 100(sample points) dimensional vector.

The resulting descriptor is then used as input to a support vector machine with a radial basis function kernel. The predictive performance is evaluated using 10x10-fold cross validation on a clonal dataset without indels relative to the 2b4c V3 region, containing 432 mutually distinct V3 regions (97 of those being X4 variants).

**Results & Conclusion:** Our novel structural descriptor considerably improves prediction of coreceptor usage. For a fixed specificity of 0.95 a sensitivity of 0.77 is achieved. This compares favorably with the sensitivity of 0.73 for purely



sequence-based prediction using a linear support vector machine. Compared to the sensitivity of 0.62 of the charge rule the increase is even more pronounced. By combining the structural descriptor with the amino acid indicator encoding of the sequence sensitivity increases further to 0.80. A detailed ROC analysis of the variability of those performance estimates reveals that the improvements are significant. In order to identify characteristic structural features for coreceptor specificity we performed a comparison of descriptor vectors for the coreceptor types

CCR5 and CXCR4. Our results indicate that modelling side-chains with SCWRL improves predictive performance compared to a simplified descriptor, where functional atoms are mapped to the C-beta atom coordinates.

Future investigations will include considering flexible backbone conformations for sequence variants and a more sophisticated handling of indels in the V3 region. Furthermore we will apply the method to sequences obtained from clinical studies, which requires the handling of sequence ambiguities.

References:

- [1] Adrian A Canutescu, Andrew A Shelenkov, and Roland L Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001-2014, Sep 2003.
- [2] Chih chin Huang, Min Tang, Mei-Yun Zhang, Shahzad Majeed, Elizabeth Montabana, Robyn L Stanfield, Dimitar S Dimitrov, Bette Korber, Joseph Sodroski, Ian A Wilson, Richard Wyatt, and Peter D Kwong. Structure of a V3-containing HIV-1 gp120 core. *Science*, 310(5750):1025-1028, Nov 2005.
- [3] Mark A Jensen and Angelique B van'tWout. Predicting HIV-1 coreceptor usage with sequence analysis. *AIDS Rev*, 5(2):104-112, 2003.
- [4] J. J. de Jong, J. Goudsmit, W. Keulen, B. Klaver, W. Krone, M. Tersmette, and A. de Ronde. Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J Virol*, 66(2):757-765, Feb 1992.
- [5] Satish Pillai, Benjamin Good, Douglas Richman, and Jacques Corbeil. A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses*, 19(2):145-149, Feb 2003.
- [6] Roland R Regoes and Sebastian Bonhoeffer. The HIV coreceptor switch: a population dynamical perspective. *Trends Microbiol*, 13(6):269-277, Jun 2005.
- [7] W. Resch, N. Hoffman, and R. Swanstrom. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology*, 288(1):51-62, Sep 2001.
- [8] Stefan Schmitt, Daniel Kuhn, and Gerhard Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, 323(2):387-406, Oct 2002.

**25 Vishal Kapoor<sup>1</sup>, Sree Nivas gurram Reddy HCL, <sup>1</sup>vishal.kapoor@hcl.in**

### ***In Silico* Pathway of drug absorption, distribution and Metabolism(ADMET)**

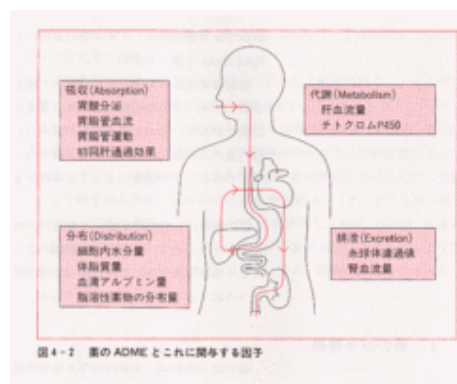


図 4-2 薬の ADMET とこれに関与する因子

**Short abstract:** Many of the drug candidates that fail in clinical trials are withdrawn because of unforeseen effects and unfavorable pharmacokinetic profiles. An ideal system would enable researchers to make a confident elimination decision based purely on the structure of a new compound and proposed in-silico parameters

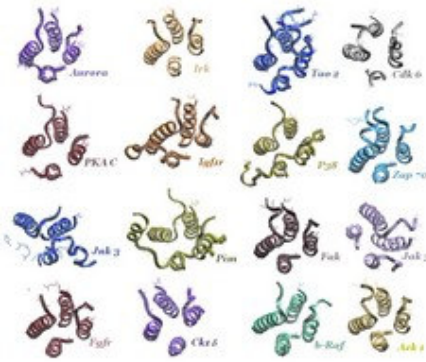
**Full abstract:** Many of the drug candidates that fail in clinical trials are withdrawn because of unforeseen effects and unfavorable pharmacokinetic profiles. An ideal system would enable researchers to make a confident elimination decision based purely on the structure of a new compound and proposed in-silico parameters ranging from data-based approaches such as QSAR & similarity searches to structure based methods such as ligand protein docking, pharmacophore calculations and homology modeling. Drug absorption and distribution parameters have been designed on the basis of mathematical equations and simulations. We classify and

summarize the major in-silico methods used to predict drug metabolism and hence, discrimination is thus made between “local” and “global” systems. So, we developed a computational algorithm for evaluating the possibility of cytochrome P450 mediated metabolic transformations that xenobiotics molecules undergo in the human body. This complete system predicts the solubility, membrane barrier, lipophilicity, volume of distribution, blood brain barrier, therapeutic index, metabolic rate & its stability, half-life, intrinsic clearance and other pharmacokinetic properties along with feasible metabolite structures of proposed drug. In short, the developed algorithm can be used in early stages of drug discovery as an efficient tool/software for the assessment of human absorption, distribution and metabolism of novel compounds in designing discovery libraries and in lead optimization.

**27 Michal Milo<sup>1</sup>, Amiram Goldblum<sup>2</sup>** The Hebrew U. of Jerusalem, <sup>1</sup>mikmillo@md.huji.ac.il

<sup>2</sup>amiram@vms.huji.ac.il

### **A Myristoyl Binding Site in Protein Kinases**



**Full abstract:** The role of a Myristoyl Binding Pocket (MBP) has been recently discovered in a single kinase, c-Abl, and shown to be the locus of autoinhibition by Abl's N-terminal myristoyl group[1]. This N-terminal myristoylation is missing in the Bcr-Abl construct, which is the reason for an elevated kinase activity that is directly linked to the induction of Chronic Myeloid Leukemia. It has therefore been suggested to design molecules that could inhibit the Bcr-Abl kinase activity by blocking the MBP, assuming that this MBP is unique to this kinase. The existence of such a MBP in other kinases has however been overlooked. We have recently discovered, by means of sequence and structure analysis and additional computational techniques, a myristoyl binding pocket (MBP) in the C-terminal lobe of the kinase domain of many families of protein kinases(see figure). In the literature, binding of myristate to Src was measured as early as 1999[2] and

was more recently suggested on the basis of an NMR experiment[3]. Direct evidence for the binding to a Src MBP is still missing, but computational studies on some 50 kinases with solved structures show that a MBP exists in all of them. We show that myristoyl binding to many of these kinases is favored by the Van der Waals component of the binding enthalpy and by a large buried surface area that provides a favorable entropic component for binding. The enthalpy of binding myristoyl to Abl is found to be the most favorable of all the kinases. Specific binding to some kinases may be possible due to variations in the residues that are in direct contact with myristate derivatives in their MBP. Many myristate derivatives have been “docked” to the MBP of several kinases and were found (in Silico) to have better binding enthalpies than for binding myristate. This is also suggested by another recent finding [4], that a molecule which is quite different than the myristoyl scaffold apparently binds to the MBP of c-Abl.

In vitro experiments of Src binding to designed myristoyl derivatives are being performed.

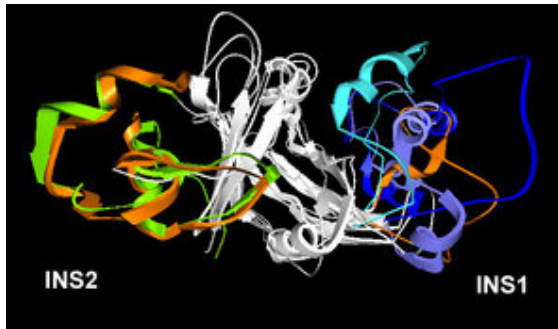
References:

1. Nagar, B. et al. Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase. *Cell* 112, 859-871 (2003).
2. Ramdas, L. et al. N-myristoylation of a Peptide Substrate for Src Converts it Into an Apparent Slow-binding Bisubstrate-type Inhibitor. *Journal of Peptide Research* 53, 569-577 (1999)
3. Cowan-Jacob SW. et al. The Crystal Structure of a c-Src Complex in an active Conformation Suggests Possible Steps in c-Src Activation. *Structure* 13, 861-871 (2005)
4. Adrian, FJ. et al., Allosteric inhibitors of Bcr-abl-dependent cell proliferation *Nature Chem. Biol.* 2, 95 – 102 (2006)

**28 Claudia Bertonati<sup>1</sup>, Marco Punta<sup>2</sup>, Burkhard Rost, Barry Honig** Columbia U., <sup>1</sup>cb2144@gmail.com <sup>2</sup>mp2215@columbia.edu

### New Perspectives from Structural Genomics - a Case Study: 5 Recently-Solved North East Structural Genomics Targets Provide New Insights into the Structure and Function of the ASCH-PUA Fold.

**Short abstract:** We analyze the structure of 5 different targets from the North East Structural Genomics consortium. All targets belong to the same fold (PUA domain-like) but share fairly low sequence similarity. Similarities and differences among the 5 targets provide new insights into their putative RNA-binding function.



**Full abstract:** The high-throughput pace at which structural genomics consortia deliver new structures is creating some new exciting opportunities. In the near future comparative structural genomics, i.e. comparative genomics powered by structural knowledge, will become a reality. Today, it is already possible to extract, from the list of solved targets, groups of proteins sharing common structural or functional features (common folds, common ligands, common metabolic pathways). Intersecting the structural information obtained from these targets with the one present in sequence, structure and function databases may provide important functional clues on previously uncharacterized proteins and biological mechanisms.

Here, we present a study on five recently solved single-domain North East Structural Genomics (NESG) targets all sharing a common fold but generally low sequence similarity. Two of them were included in a recent study by Aravind and coworkers, in which the authors used structural comparisons and sensitive sequence alignment methods to identify a new phyletic lineage that they named the ASCH-family [1]. In the same study, they uncovered a structural link with the so-called PUA fold, and suggested an RNA-related function for the ASCH-family. The three newly solved NESG targets, one of which was crystallized in complex with three small ligands, allow more stringent testing for these early hypotheses. We perform structural comparison, electrostatic potential calculations, sequence conservation analysis and genomic co-localization studies on all 5 NESG targets. The new data strengthen the idea that ASCH domains bind RNA, at the same time suggesting higher diversity than previously anticipated in the mode



of binding. Also, comparison with double-domain PUA proteins seems to indicate that single ASCH-PUA domains may bind RNA by acting in concert with other proteins, rather than in isolation.

[1] Iyer LM, Burroughs AM, and Aravind L. *Bioinformatics*. 2006; 22(3):257-63. The ASCH superfamily: novel domains with a fold related to the PUA domain and a potential role in RNA metabolism.

**31 Peter Røgen<sup>1</sup>, Per W Karlsson<sup>2</sup>** <sup>1</sup>Department of Mathematics, Technical U. of Denmark, Peter.Roegen@mat.dtu.dk  
<sup>2</sup>P.W.Karlsson@mat.dtu.dk

### Quantifying the Relative Positions of Proteins Secondary Structure Elements



**Short abstract:** How parallel and anti-parallel are the helices of your favourite protein? - And which other proteins have a similar parallelity of their helices? To answer such questions, we introduce a notion of Local Self Parallelity of protein secondary structures, and use it for automatic classification of protein structures.

**Full abstract:** How parallel and anti-parallel are the helices of your favourite protein? - And which other proteins have a similar parallelity of their helices?

We quantify the contents and the relative positions of the secondary structure elements of a protein structure in a way that is continuous and smooth under deformation of the protein structure. To do this we introduce:

- 1) a notion of Local Self Parallelity of a space curve and
- 2) a real-valued smooth (alpha,beta,coil)-coloring of the backbone giving

3) the alpha-alpha, the alpha-beta, the alpha-coil, the beta-alpha etc.

Local Self Parallelity of a protein chain.

Hereby we can quantify how parallel the helices and the sheets of a protein are, or e.g. how parallel the sheets and the "coil"-regions of the protein are, etc. Using the N-to-C direction of the backbone, we furthermore distinguish between the helices lying upstream and downstream relatively to the sheets. That is: To the alpha-beta Local Self Parallelity only the helices lying upstream relatively to the sheets contribute and to the beta-alpha Local Self Parallelity only the helices lying downstream relatively to the sheets contribute.

The algorithm starts with the carbon alpha curve given by a pdb-file. This curve is then smoothened by a local linear filter [1,2] and the (alpha,beta,coil)-coloring is given as a spline-function of neighbour and next-neighbour distances on the smoothened backbone. Using the (alpha,beta,coil)-coloring as weights the measures of Local Self Parallelity are linear maps from the distance matrix of the smoothened backbone. The calculation time is thus very short (> 1000 domains per minute including reading the pdb-files).

As functions of the distance matrix, these structural measures are blind to chirality. It is thus a bit surprising that a Jack knife test shows that they can automatically classify 96% of 18861 connected CATH-domains correctly which is the same rate of success as the highly chirality sensitive Gauss integrals was reported to have in [3]. However the error rate of 1% here is a lot higher than the Gauss integrals 0.02% error rate reflecting that eg. four helix bundles come in a right-handed and in a left-handed tertiary structure which is indistinguishable by these distance matrix based structural measures.

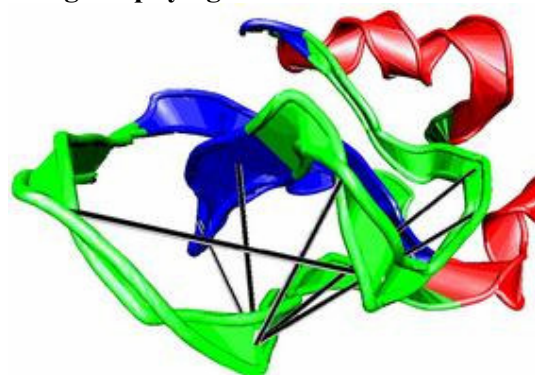
[1] P. Røgen, Evaluating protein structure descriptors and tuning Gauss integral based descriptors *J. Phys.: Condens. Matter* 17 (2005) S1523-S1538928.

[2] K. Lindorff-Larsen, P. Røgen, E. Paci, M Vendruscolo & C.M. Dobson, Protein folding and the organization of the protein topology universe, *Trends in Biochemical Sciences*, 30(1), 13-19. 2005.

[3] P. Røgen & B. Fain, Automatic classification of protein structure by using Gauss integrals, *PNAS*, 100(1), 119-124, 2003.

**34 Scooter Willis<sup>1</sup>** <sup>1</sup>U. of Florida-Department of Computer and Information Science and Engineering, willishf@ufl.edu

### Predicting co-evolving pairs in Pfam based on Mutual Information of mutation events measured along the phylogenetic tree



**Short abstract:** Mutual Information (MI) is used to detect co-evolving pairs in a protein family by re-sampling based on mutation events in the phylogenetic tree. The predictive quality with MI ( $Z \geq 4$ ), a sequence distance > 10 and within 12 angstroms using the RPE method is on average 81% in a Pfam family.

**Full abstract:** The accurate prediction of co-evolving pairs in protein sequences plays an important role in tertiary protein structure prediction and protein engineering. The use of mutual information or entropy measures to detect co-evolving pairs is an actively researched topic [1],[2],[3]. The phylogenetic effect on sequence data in a protein family impacts accurate probability measurements when calculating entropy for a sequence position. In this paper, a method is introduced to reduce the phylogenetic effect (RPE method) on probability calculations used to determine Mutual

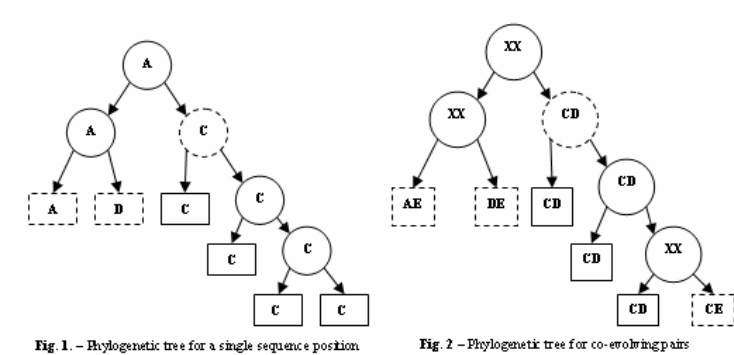
Information (MI) between two sequence positions. Two sequence positions that have a MI score that is equal to or greater



than four standard deviations from the mean ( $Z \geq 4$ ) for that family are considered statistically significant and potential co-evolving pairs or highly correlated. A co-evolving pair can be defined that when one amino acid mutates a neighboring amino acid in 3D space may also mutate to preserve protein structure or function. The predictive quality of co-evolving pairs that have a sequence distance  $> 10$ , with high mutual information ( $Z \geq 4$ ) and are less than 12 angstroms apart using the RPE method is on average 81% in a Pfam family. The accuracy of detecting co-evolving pairs in Pfam using MI, where  $Z \geq 4$  and standard probability measurements is 56% (STA method). This study represents the first known analysis of MI to detect co-evolving pairs in the full Pfam data set. Results for each protein family in Pfam and corresponding ribbon models graphing the MI relationships can be found at [www.protein3d.com](http://www.protein3d.com).

Information Theory Approach: Mutual information is based on Shannon Entropy (H) which is derived from the probabilities of occurrences of individual and combined events between two discrete random variables. Using a phylogenetic tree for a protein family it is possible to detect mutation events at a particular sequence position. This reduced set of mutations could then be used as the basis for probability calculations to calculate the entropy for that sequence position. In Figure 1, a binary tree represents a hypothetical phylogenetic tree where the circles are the theoretical parent and the boxes represent a sequence position in a protein family. Without taking into consideration the phylogenetic influence the probability of the sequence position is  $p(A)=1/6$ ,  $p(D)=1/6$  and  $p(C)=2/3$ . The RPE method calculates probability of the sequence position by starting at the root node and counting all children nodes where the child does not equal the parent (represented by the dashed lines in Figure 1) resulting in probability calculations of  $p(A)=1/3$ ,  $p(D)=1/3$  and the  $p(C)=1/3$ . The first approach accurately represents the population sample and the latter represents the probability of transition to a different amino acid.

In Figure 2, the tree represents a pair of amino acids found at sequence positions x and y. The phylogenetic tree is used to detect mutation events between pairs which becomes the population sample used for probability calculations. The probability based on the number of observed pairs between two sequence positions would result in  $p(AE)=1/6$ ,  $p(DE)=1/6$ ,  $p(CD)=3/6$  and  $p(CE)=1/6$ . By using the method described above we start at the root node and count children nodes (represented by the dashed lines) that are different with the additional rule that if an internal node is XX it takes on the value of its parent node and is not counted, we get the following probabilities:  $p(AE)=1/4$ ,  $p(DE)=1/4$ ,  $p(CD)=1/4$  and  $p(CE)=1/4$ . Using the STA method the CD sequence pair occurs 50% of the time but using the RPE method it is reduced to 25%.



Mutual Information in Protein Families  
Once mutual information is calculated for a protein family the referenced PDB models for that family are used to determine actual 3D position/distance between amino acid pairs as the closest non-Hydrogen atoms. The primary focus is on mutual information scores with a value four times or greater than the standard deviation from the mean or  $Z \geq 4$  and sequence distance between pairs greater than 10. The average prediction percentage score for each method represents the likelihood that if we use the same approach in Pfam families that do not have solved PDB models that the

MI scores would represent co-evolving pairs. The prediction accuracy of MI to detect co-evolving pairs in each protein family is determined and then averaged for all families to determine the overall prediction accuracy of each method. To improve the prediction accuracy, various data attributes of the measurements are used as cutoffs to eliminate false positives. By including an additional filter or constraint on the overall number of predicted co-evolving pairs in a family where the number of pairs with  $Z \geq 2$  is less than 500 increases the  $Z \geq 4$  percentages to 56.2% for the STA method. For the RPE method it was determined that if the number of mutation events between co-evolving pairs was less than 40 the prediction accuracy was low. For the RPE method the additional filter is applied where the number of mutation events for a sequence pair must be  $> 40$  and #MI pairs  $< 500$  results in overall improved prediction accuracy of 81.3%. Using the filter criteria, the accuracy of the STA and RPE methods are compared in Table 1.

Table 1- #MI scores per family  $< 500$  and RPE MC  $> 40$

	STA		RPE	
	% $<12A$	%Pfam	% $<12A$	%Pfam
$Z \geq 4$	56.2	18.3	81.3	15.8
$Z=3$	42.6	55.8	56.4	46.4
$Z=2$	33.2	79.7	36.9	71.6

Conclusion: The RPE method of re-sampling sequence data based on mutation events along the phylogenetic tree is an effective approach in improving the quality of predicted co-evolving pairs. Future work will focus on using predicted co-evolving pairs using the RPE method in a protein family for tertiary protein prediction and detecting potential regions of structural or functional importance.

[1]Crooks,G., Wolfe,J., and Brenner S. (2004), Measurements of

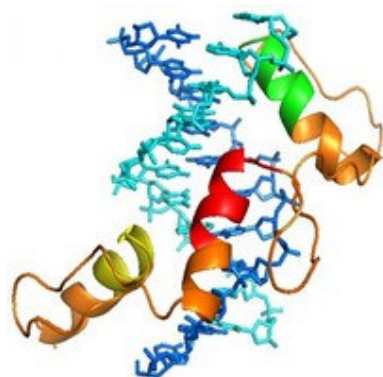
Protein Sequence Structure Correlations, PROTEINS: Structure, Function, and Bioinformatics, 57, 804-810

[2]Atchley,W.R. et al. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol. Evol.*, 17, 164-178.

[3]Hamilton,N., Burrage,K., Ragan,M.A. and Huber,T. (2004) Protein contact prediction using patterns of correlation. *Proteins*, 56, 679-684.

**39 Jeff D Sander<sup>1</sup>, Peter C Zaback<sup>2</sup>, Drena L Dobbs<sup>3</sup>, Fengli Fu, Daniel F. Voytas** Iowa State U. Bioinformatics and Computational Biology, <sup>1</sup>jdsander@iastate.edu <sup>2</sup>petez@iastate.edu <sup>3</sup>ddobbs@iastate.edu

### Designing C2H2 Zinc Finger Proteins to Target Specific Genomic Sites



**Short abstract:** Consisting of modular nucleic acid binding domains, C2H2 zinc finger proteins provide an excellent framework for engineering new sequence-specific DNA binding proteins. Using binding specificities determined by others, we have developed a program, ZiFit, to locate and rank candidate targets for zinc finger binding within a given DNA sequence.

**Full abstract:** Zinc fingers, the most abundant DNA binding motifs in eukaryotes, provide one of the best understood protein-DNA binding mechanisms. They promise to become valuable tools for genome modification and clinical intervention in disease because they can be used to target proteins, including nucleases and transcription factors, to virtually any desired location in any genome. Consisting of multiple modular and interchangeable nucleic acid binding domains, C2H2 zinc finger proteins provide an excellent framework for engineering new sequence-specific DNA binding proteins (see [1]).

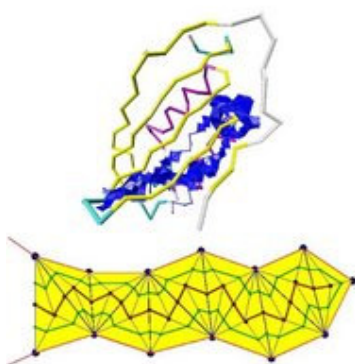
The canonical C2H2 zinc finger protein comprises three individual fingers. Each finger consists of a beta sheet and an alpha helix coordinated through a zinc ion via two cysteine and two histidine residues. Positioned within the major groove, the helices form both base-specific and backbone contacts with adjacent nucleotide triplets (Figure 1A). Amino acids -1, +3, +6 (with respect to the start of the  $\alpha$ -helix (Figure 1B)) each make base-specific contacts with a single nucleotide on the 5' to 3' DNA strand. In addition, the amino acid at position +2 can contact a base preceding the triplet on the 3' to 5' strand.

Using C2H2 zinc finger binding specificities determined by others (e.g. [2], [3]), we have developed a program, ZiFit (see [4]) to locate and rank candidate targets for zinc finger binding within a given DNA sequence (Figure 2). In ongoing work, we are designing and experimentally testing additional zinc finger binding modules and combinations by exploiting knowledge-based approaches that incorporate information such as binding affinities, context dependence, and DNA sequence/structural characteristics.

In recent work we have begun to explore molecular dynamics simulations to enhance the design of zinc fingers. By making specific substitutions in either the DNA or protein component of existing zinc finger-DNA co-crystal structures, we can explore the space of zinc finger-DNA complexes to computationally estimate the relative binding free energies and specificities of zinc finger proteins.

**40 Bala Krishnamoorthy<sup>1</sup>** <sup>1</sup>Washington State U., kbala@wsu.edu

### Characterization of Protein Structure using Geometry and Topology



**Short abstract:** We define the neighborhood of a contiguous strand of amino acids in a protein as a series sub-complexes of the alpha complex of the protein at various growth levels. We characterize units of protein structure (motifs) by analyzing the geometry and topology of these neighborhood complexes.

**Full abstract:** Motivation: There is a need for an objective, and at the same time, automated characterization of protein structure at all levels - from secondary structure to classification of proteins into structural families. Popular structural databases such as SCOP and CATH involve visual inspection of structures as one of the key steps, while automated methods (such as knot theory-based techniques) overlook the characterization of parts of protein structure.

Results: We propose a novel framework for the characterization of all levels of protein structure based entirely on the geometry of its parts. The three-dimensional t-complex filtration (commonly referred

to as the alpha complex) of the protein represented as a union of balls (one per residue) captures all the relevant information about its geometry and topology. The neighborhood of a strand of contiguous alpha carbon atoms is defined as a retracted "tube" which is a sub-complex of the original complex.

Imagine a tube that has the back-bone chain running through its center. If we thicken the tube uniformly, the parts of the tube around the strand in question would come into contact with the surface of the tube around other parts depending on the interactions between the strand and the other parts of the protein. The tube would touch itself if the strand bends on itself, thus allowing us to characterize domains of this type. At the same time, when this tube touches itself, one or more simplexes (vertices, edges, or triangles) in the tube complex get identified with certain others. It is desirable if we could actually create a copy of any such simplex, and pull the copy just away from the original simplex, such that tube is not self-

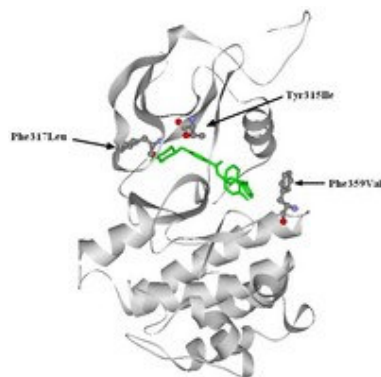
intersecting any more. The critical point in performing such a duplication is that we do not desire to alter the (topological) connectivity of the original tube in this process. For this purpose, we retract the original tube closer towards the back-bone while preserving its homology. We capture the topology of the retracted tube by computing the most persistent connected components and holes in its filtration.

A "motif" for 3D structure is characterized by the number of persistent components and holes, and their relative persistences in the filtration of the tube complex. The final retraction step helps to simplify the tube complex topology, and allows the persistence calculations to be performed in near-linear time.

We describe the salient features of these motifs, and suggest various uses of the same, which are currently being investigated. For instance, we took a set containing ten chains each from the SCOP classes a (all alpha), b (all beta), c (alpha/beta), and d (alpha + beta). A cluster analysis using the counts of these motifs was able to successfully group them into separate hierarchies quite efficiently. Several similar samples (of 40 chains with 10 each from the classes a,b,c,d) could be classified with an average accuracy of 83%.

**41 Sergio A Alencar<sup>1</sup>, Julio C Lopes, Andrelly M. Jos<sup>1</sup>** <sup>1</sup>Departamento de Quimica, Universidade Federal de Minas Gerais, sergiodealencar@bioinfo.dout.ufmg.br

## Chemoinformatics approach to Differential Drug Response of the Tyrosine Kinase Domain BCR-ABL to Imatinib in Chronic Myeloid Leukemia Patients



**Short abstract:** Chemoinformatics approach to differential drug response of the tyrosine kinase domain BCR-ABL to imatinib in chronic myeloid leukemia (CML) patients in order to investigate the effects of three substitutions (Tyr315Ile, Phe317Leu and Phe359Val) at direct drug contact sites in CML patients that develop imatinib resistance.

**Full abstract:** The tyrosine kinase domain BCR-ABL is the fusion product of a reciprocal chromosome translocation between chromosomes 9 and 22, known as the Philadelphia (Ph) chromosome, and is present in the leukemic cells of more than 95% of patients with chronic myeloid leukemia (CML) (1). The tyrosine kinase inhibitor STI571 (imatinib/gleevec) is currently used in the treatment of CML patients. Differential drug response to this inhibitor in patients with CML has been associated with tyrosine kinase domain mutations (2) Fifteen different amino acid substitutions affecting 13 residues in the kinase domain have been identified in 29 out of 32 patients whose disease relapsed after an initial response to imatinib

(3). Through computational analysis we have investigated the effects of three substitutions (Tyr315Ile, Phe317Leu and Phe359Val) at direct drug contact sites in CML patients that develop imatinib resistance.

The mouse (*Mus musculus*) tyrosine kinase domain structure in complex with imatinib (PDB ID: 1IEP) was obtained from the Protein Data Bank (PDB). The amino acid substitutions were introduced in the appropriate positions in the protein structure using the mutate tool of Deep View. The receptor geometry was optimized using the TINKER (4) package with the following parameters: AMBER99 force field, dielectric coefficient: 4.0, nobond cutoff: 14.0, and the steepest descent convergence method. Hence the most energetically favorable conformations for each structure were obtained.

The ligand (imatinib) was constructed using the program Ghemical (5), and its charges calculated with the AM1 (MOPAC 7.0) semi-empirical method through the ESP procedure.

Through the implementation of a genetic algorithm (GA) based on mechanisms of biological evolution, AutoDock 3.0 (6) allows efficient ligand positioning at the active site of the protein target. Rotation was considered between the bonds of carbon and the hydroxyl oxygen, and between carbons of the aromatic ring. In order to use the AutoDockTools program, hydrogens, Kollman atomic charges and solvent parameters were added. Hence AutoDock was used to predict how imatinib binds to the tyrosine kinase domain receptor, allowing comparisons between the mutants and the wild type receptor. AutoDock analysis was carried out with each mutant and imatinib, showing free energy of binding values greater than that obtained with the wild type (Table 1). Previous work on the biochemical characterization of imatinib resistance (2) has agreed with our results for the affinity between target and drug (Table 2).

Single nucleotide polymorphism (SNP) studies identify amino acid substitutions in protein-coding regions. Each has the potential to affect protein function (7). Based on multiple alignment information, SIFT (Sorting Intolerant From Tolerant) predicts whether an amino acid substitution affects protein function, distinguishing between functionally neutral and deleterious amino acid changes on human polymorphisms. SIFT returns predictions on whether the substitutions are tolerant or intolerant based on the scores (8). SIFT predicted that two of the substitutions (Tyr315Ile and Phe359Val), which occur at highly conserved positions across species, are intolerant and therefore have a phenotypic effect (Table 1).

Table 1 TKD AutoDock GA (Kcal/mol) SIFT Score

Wild	-7.71	-
Tyr315Ile	-2.86	intolerant (0.05)
Phe317Leu	-7.16	tolerant (0.62)
Phe359Val	-4.06	intolerant (0.00)

TKD: Tyrosine Kinase Domain

Table 2

TKD Biochemical IC50 (mM)

Wild	0.28
Tyr315Ile	> 10
Phe317Leu	3.30
Phe359Val	Not assessed

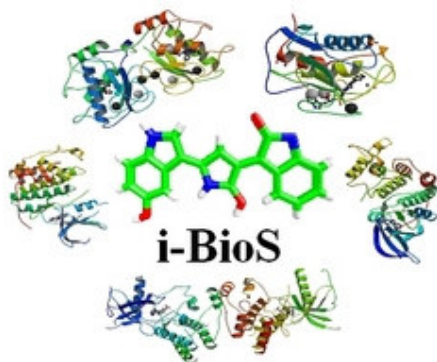
The results of our work indicate that the substitutions studied affect drug affinity at varying levels, suggesting the need for drug alternatives in the treatment of imatinib-resistant CML patients. Our results also confirm previous imatinib biochemical assays, showing varying degrees of resistance to imatinib caused by mutations in the receptor site.

#### References

- 1 - Wong S and Witte ON (2004) The BCR-ABL Story: Bench to Bedside and Back. *Annu. Rev. Immunol.* 22:247-306.
- 2 - Shah NP, Nicoll JM, Nagar B, Gorre ME, Paquette RL, Kuriyan J and Sawyers CL (2002) Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell.* 2:117-125.
- 3 - Gorre ME, Mohammed M, Ellwood K, Hsu N, Paquette R, Rao PN and Sawyers CL (2001) Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science.* 293(5531):876-80.
- 4 - <http://dasher.wustl.edu/tinker>
- 5 - <http://www.uiowa.edu/~gchemical>
- 6 - Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK and Olson AJ (1998) Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. *J. Computational Chemistry.* 19:1639-1662.
- 7 - Ng PC and Henikoff S (2001) Predicting deleterious amino acid substitutions. *Genome Res.* 11(5):863-74.
- 8 - Ng PC and Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 12(3):436-46.

**42 Julio CD Lopes**<sup>1</sup> Cheminformatics and Medicinal Chemistry Group - Departamento de Química – UFMG, [jlopes@netuno.lcc.ufmg.br](mailto:jlopes@netuno.lcc.ufmg.br)

### **i-BioS: A System for Inverse Biological Virtual Screening Based on Inverse Docking Approach**



**Short abstract:** In the present work we develop a new system for inverse docking studies. Our system, called i-BioS (Inverse Biological virtual Screening), is based on Surflex docking program was used to reproduce some of the biological activities of violacein and found new ones, associated to its anti-oxidant character.

**Full abstract:** In a competitive environment such as the pharmaceutical industry, getting first on the market is vital. Despite the long and massive investments on promising technologies such as HTS and combinatorial chemistry, the cost of bringing a successful drug to market is estimated between US\$900 million (1) and US\$1.7 billion (2) per single successful drug marketed. These loads on the costs of all the failures, around 60-70% can be attributed to failures during the various stages of drug discovery and development. One of the major bottle neck is the selection of lead candidates which will prove successful in pre-clinical

and clinical development. Within drug discovery, most failed compounds have problems associated with their ADME or toxicity profile.

The great demand on improving the efficiency of drug discovery has created a need for a new paradigm that enables the use structural information in the combinatorial chemistry and medicinal chemistry process. A variety of tools have been developed to identify lead candidates in extensive collections of compounds, like virtual screening and high throughput docking. Additionally, it is becoming clear that successful prediction of drug-like properties at onset of drug discovery will payoff later in drug development (3).

Recently, an innovative approach called inverse docking had been introduced as a chemoinformatics tool to predict a priori several characteristics of lead candidates. (4,5). From a database of biological and therapeutic targets, one could (i) identify unknown and secondary therapeutic targets for a drug, (ii) predict the potential toxicity and side effects of an investigating drug, and (iii) probe the molecular mechanisms of action of bioactive compounds.

As the docking technique has reached the position of established technique, the crucial step for applying the inverse docking is the correct selection of the biological and therapeutics targets. Therapeutic effects of a drug generally result from its interaction with one or more proteins or nucleic acids critical in a disease process, the therapeutic targets. The adverse reaction of a drug in the human body is often induced by its interaction with some of the proteins critically important. Identification of the interaction with these proteins, known as toxicity and side effect targets, has potential application in facilitating the prediction of side effects and toxicity of an investigative drug in early stage of drug development.

The inverse docking approach has been applied by a few groups in the world in order to make a virtual screening of biological and pharmacological properties of molecules of interest. The Bioinformatics & Drug Design group (BIDD) (6) has developed the INVDOCK software. (4) The Laboratoire de Bioinformatique du médicament group (7) developed the sc-PDB, (5) a database of targets for inverse docking studies, comprising a collection of more than 4000 active sites from the PDB.

In the present work we use all protein structures in sc-PDB to develop a new system for inverse docking studies. Our system, called i-BioS (Inverse Biological virtual Screening), is based on Surflex docking program (8) from Biopharmics



LLC (9). After obtaining the targets collection, the Reduce program (10) was used for adding all hydrogens atoms with optimization of the orientations of polar hydrogens. Subsequently, each target was submitted to Surflex in order to generate the protomol representation of actives sites. The position of the binding site was determined by coordinates of ligand molecules in the sc-PDB.

The protomol files were archived and we have designed a script which connect the protomol files, the Surflex docking module and the OpenEye Omega software (11). Omega is used to generate 3D conformation structure of the query molecule. The user can input a molecule in the system in a variety of formats such as smiles, SDF or mol2. After the conformer generation the script calls the Surflex program and calculates the geometry and docking energy for the complexes of input molecule with all proteins in the database. The output is a rank of the proteins, with the most probable targets at the top.

We have applied i-BioS system to study the potential targets of violacein, a pigment isolate from *Chromobacterium Violaceum*, a bacterium found in the Black River, an affluent of the Amazon River. Violacein presents a variety of biological effects, such as anti-cancer, anti-protozoa and anti-viral activities. Using our i-BioS system we were able to reproduce some of these biological activities and found new ones, mainly those associated to its anti-oxidant character. Other results will be discussed concerning applications of i-BioS to guide the biological screening of a selection of natural products isolated by phytochemistry groups from Brazil.

#### References

- 1 - Anon (2003) Total Cost to Develop a New Prescripton Drug, Including Cost of Post-Approval Research, is \$897 Million," The Tufts Center for the Study of Drug Development, Press Release, May 13, 2003.
- 2 - R. Mullin (2003) Drug Development Costs About \$1.7 Billion Chem Eng News, 81:8, 2003.
- 3 - Lipinsky CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46:3-26, 2001
- 4 - Chen YZ, Li ZR, Ung CY (2002) Computational Method for Drug Target Search and Application in Drug Discovery. *J. Theor. Comp. Chem.*, 1, 213-224. (2002).
- 5 - Paul N, Kellenberger E, Breat G, Muller P, Rognan D (2004) Recovering the True Targets of Specific Ligands by Virtual Screening of the Protein data Bank Proteins, 54:671-680, 2004
- 6 - <http://bidd.nus.edu.sg/group/bidd.htm>
- 7 - <http://bioinfo-pharma.u-strasbg.fr/bioinformatics-cheminformatics-group.php>
- 8 - Jain AN (2003) Surflex: Fully Automatic Flexible Molecular Docking using a Molecular Similarity-Based Search Engine. *J Med Chem* 46, 499-511, 2003.
- 9 - <http://www.biopharmics.com/>
- 10 - <http://kinemage.biochem.duke.edu/software/reduce.php>
- 11 - <http://www.eyesopen.com/products/applications/omega.html>
- 12 - Brazilian National Genome Project Consortium (2003) The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability PNAS, 100: 11660-11665, 2003

**47 Prashanth k. Aitha<sup>1</sup>, Zaharul Ahsan<sup>2</sup>, Srinivasa Reddy Gurram** HCL Technologies,

<sup>1</sup>prashanth.aitha@hcl.in <sup>2</sup>mahsan@hcl.in

#### Toxicity Predication from Chemical Structure as a process of Lead optimization.



**Short abstract:** A model to predict the toxicity of drug like molecules based on its structural similarity to existing drugs has been proposed using Structure activity relationship (SAR) and other Statistical approaches.

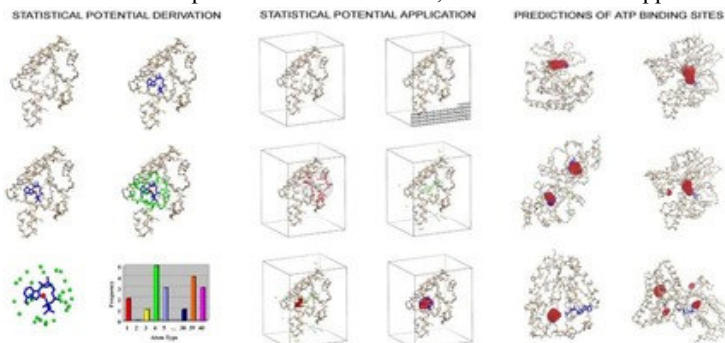
**Full abstract:** The Prediction of toxicity prior to actual experimentation from chemical structure can make a valuable contribution to the reduction of animal usage in the screening out of potentially toxic chemicals at an early stage and in providing data for making positive classifications of toxicity. Properties like toxicity are implicit in its chemical structure. Quantitative modeling methods, relating aspects of chemical structure to biological activity, have long been applied to the prediction and characterization of chemical toxicity. Starting from the early linear free-energy approaches of Hansch and Free Wilson to correlate chemical structure with biological

activity different models were developed to suitably predict the toxicity. All the models can be classified into two different approaches namely "Expert Rule Based System" and "Structure Activity Relationship (SAR) based system". SAR when combined with computational Neural Networks (CNN) has been proved to be the effective for classifying a new drug molecule to be toxic or non-toxic. Hence a suitable model is designed to predict the toxic nature of lead compounds. The model would help us to optimize the lead compounds and simultaneously reduce the time and wastage of resources being spent upon the drug discovery process and hence to optimize the lead compounds.

**54** Francisco Melo<sup>1</sup>, Evandro Ferrada<sup>2</sup> Catholic U. of Chile, <sup>1</sup>fmelo@bio.puc.cl <sup>2</sup>evandro.ferrada@gmail.com

## Implicit Statistical Potentials for the Characterization and Prediction of Protein-ligand Binding Sites

**Short abstract:** Here we will describe the first known method for protein-ligand binding site prediction that uses statistical potentials derived from experimental data and does not depend on explicit coordinates of a ligand nor transformations in three-dimensional space. The method is fast, robust and can be applied to any ligand.



**Full abstract:** The ongoing structural genomics project is generating a large amount of new protein fold structures that are deposited daily at the Protein Data Bank (PDB). Additionally, several large-scale protein structure modeling efforts are using the available experimental data to produce millions of three-dimensional protein structure models, which constitute a rich source of information to infer and annotate protein function more accurately than from sequence information alone. The reason of this is that most of

specific protein-ligand binding sites and protein function in general depend directly on the three-dimensional structure of a protein rather than on its primary sequence of amino acids. However, due to the lack of sufficient knowledge about the structure and function relationship in proteins, the development of novel and fast methods is required in order to cope with the analysis of the vast amount of structural information that is currently being generated.

A new method based on statistical potentials, which does not depend on the explicit coordinates of the ligand atoms, has been developed to describe and predict any protein-ligand binding site at the atomic level. The method is generic and robust, thus it can be applied to any small protein ligand for which enough experimental data is available. The method is validated using all known ATP binding sites from the experimental ATP-protein complexes currently available at the PDB. The sensitivity and specificity trade off can be easily adapted depending on the particular application. The method is fast enough to be applied in a large-scale, provided that a high specificity threshold is used and that sufficient computing resources are available. It is suggested that this method could be used in a large-scale to bias the selection of putative protein-ligand binding sites to be assessed by the more accurate but CPU expensive docking simulations that depend on the explicit atomic coordinates of the protein-ligand complex.

---

**60** Oxana V Galzitskaya<sup>1</sup>, Michail Yu Lobanov<sup>2</sup>, Sergiy O Garbuzynskiy Institute of Protein Research, <sup>1</sup>ogalzit@vega.portres.ru <sup>2</sup>mlobanov@phys.portres.ru

## Prediction of Amyloidogenic and Unfolded Regions in Protein Chains

**Short abstract:** Identification of potentially amyloidogenic and unfolded regions in polypeptide chains is very important because such regions are essential for protein function.

**Full abstract:** Identification of potentially amyloidogenic and unfolded regions in polypeptide chains is very important because such regions are essential for protein function. In our work we introduce a new parameter, namely mean packing of residues (packing density) to detect amyloidogenic and unfolded regions in a protein sequence. It has been shown that regions with strong expected packing of residues would be responsible for the amyloid formation. Our predictions are consistent with known disease-related amyloidogenic regions for 8 of 11 amyloid-forming proteins and peptides in which the positions of amyloidogenic regions have been revealed experimentally. Our findings support the concept that the mechanism of amyloid fibril formation is similar for different peptides and proteins. Moreover, we have demonstrated that regions with weak expected packing of residues would be responsible for the appearance of unfolded regions. Our method (FoldUnfold) has been tested on datasets of globular proteins (559 proteins) and long disordered protein segments (129 proteins) and showed improved performance over some other widely used methods, such as DISOPRED, PONDR VL3H, IUPred, GlobPlot.

**61** youssef wazady<sup>1</sup> <sup>1</sup>ESTC, wazady@yahoo.fr

## Conformational Analysis of N-Acetyl-N' Methylecyclopentyl Amide

**Short abstract:** the present study is interested in analysis of the structural behaviors of Acc5. The method used in this study is based on principles of the classical mechanics. It is applied with the PEPSEA program. The use of this residue in chemotactic peptides is discussed.

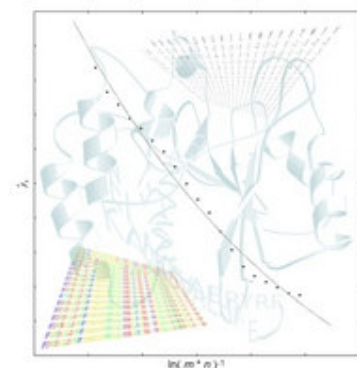
**Full abstract:** In recent years, considerable interest has been focused on the chemotactic peptide formyl-Met-Leu-Phe-OH "fMLPOH" which has been shown to induce lysosomal enzyme release that plays a very important role in the immunological system. Since it was reputed to be a very active agent, various studies have been realized in order to better understand this tripeptide. Among these studies we can quote the substitution of the Leu residue by the -disubstituted amino acids such as Acc5 (1- aminocyclopentane-1-carboxylic acid). This residue was the subject of a considerable number of conformational -disubstituted carbon atom in this amino acid studies. The presence of a confers upon it the

property to have a considerable steric effect, which can then impose significant constraints on the orientation of the peptide backbone. The use of this residue in a given peptide will facilitate the determination of the conformation adopted during the interaction with the receptor, as well as the search of pharmacophore groups.

In the framework of the structure–activity relationship studies, and in order to explore the conformational space of chemotactic peptides, the present study is interested in analysis of the structural behaviors of Acc5. The method used in this study is based on principles of the classical mechanics. It is applied with the PEPSEA program.

In a first step, we have calculated and integrated the energy and geometrical parameters of the Acc5 in the force field of the method. Then we have realized the conformational study of the mono-peptide N-acetyl-N'-methylcyclopentyl amide (Ac-Acc5-NHMe). Obtained results are compared to those of the recent studies. The use of this residue in chemotactic peptides is discussed.

**65 Mindaugas Margelevicius<sup>1</sup>, Ceslovas Venclovas<sup>2</sup>** Institute of Biotechnology, <sup>1</sup>minmar@ibt.lt <sup>2</sup>venclovas@ibt.lt  
**A New Approach for Estimation of Statistical Significance in Sequence Profile-Profile Comparisons**



**Short abstract:** Sequence profile-profile comparison is one of the most sensitive techniques for distant homology detection. We propose a new approach to estimate statistical significance of the profile-profile alignments. The exhaustive testing showed that this approach combined with a newly developed comparison procedure compares favorably to a number of other existing methods.

**Full abstract:** The detection of distant relationship between proteins is a common approach to structural/functional annotation of uncharacterized proteins. Currently, sequence profile-to-profile comparison is one of the most sensitive techniques for distant homology detection. However, the sensitivity of particular profile-profile comparison method is strongly dependent on the efficiency of statistical significance estimation. In this study, we propose a new approach to estimate statistical significance of the profile-profile alignments. As in the case of other existing methods, the parameters for this method were derived from the alignment score distributions. However, in contrast to many approaches that use random shuffling

for obtaining unrelated sequences, we have used real biological sequences from ASTRAL database. All-to-all alignments were derived using newly developed profile-profile comparison tool. The alignments then were clustered according to the length of profiles, and alignment score distributions were drawn individually for each cluster. The results clearly showed the dependence of the statistical parameters on the profile lengths. The formal expressions were found to be different from those used in profile comparison methods such as COMPASS. As a result, new mathematical expressions describing best distributions of the statistical significance parameters were derived. These significance estimates together with the newly developed profile-profile comparison tool showed an increased ability to detect links between distantly related proteins. The newly developed approach has been found to produce significantly better overall results than either PSI-BLAST, a profile-sequence matching tool, or PROF\_SIM, a sensitive profile-profile comparison method.

**68 Fabrice David<sup>1</sup>, Anne-Lise Veuthey<sup>2</sup>, Yum Lina Yip<sup>3</sup>** Swiss Institute of Bioinformatics - Swiss-Prot group, <sup>1</sup>fabrice.david@isb-sib.ch <sup>2</sup>anne-lise.veuthey@isb-sib.ch <sup>3</sup>lina.yip@isb-sib.ch

### **Incorporation of 3D Structure Information Into Swiss-Prot Annotation Workflow**

**Short abstract:** In this work, we created an integrated platform to obtain consistent information on protein structures, and to semi-automatically generate UniProtKB/Swiss-Prot annotation from structural features (e.g. ligand binding sites). For this purpose, we set up a structure-sequence mapping. This mapping was already used to semi-automatically annotate disulfide bonds in UniProtKB/Swiss-Prot.

**Full abstract:** The UniProtKB/Swiss-Prot knowledgebase is a comprehensive and curated protein sequence database. Swiss-Prot annotators collect and summarize current knowledge from literature. Even though structural information can provide additional evidence for certain protein features, the use of this information is not yet well integrated into the annotation workflow. Annotators may have to consult several external resources in order to get an overview of related structural information or even to inspect the raw data from PDB files. The aim of this project is to create a tool that provides annotation from structural features. The first step towards this goal is to create a reliable and easily exploitable structure-sequence mapping. Currently, a mapping is provided by the MSD (Macromolecular Structural Database) group to generate PDB cross-references in UniProtKB entries. However, this mapping is taxonomy-specific and authorizes a PDB sequence to be associated to only one Swiss-Prot entry (except for fusion proteins). We developed an automatic mapping procedure to help link interesting structural data to UniProtKB/Swiss-Prot sequences without these constraints on taxonomy and sequence identity. In addition, isoform sequences are taken into account in our pipeline, which is not yet the case in the reference mapping provided by MSD.

Key steps of the mapping are the reconstruction of sequences from PDB structures, and the production of local alignments between these reconstructed sequences and UniProt sequences. PDB sequences were reconstructed mainly on the basis of atomic coordinates. Local alignments were obtained through BLAST searches, and only hits with a sequence identity greater than 50% (85% for UniProtKB/TrEMBL entries) were kept. Both steps were designed to allow a weekly

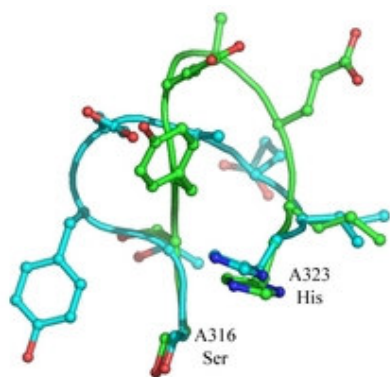
incremental update of the mapping, which is essential to provide up-to-date information to annotators. PDB data, reconstructed sequences and alignments are stored into a PostgreSQL relational database to allow a fast on the fly structure-to-sequence or sequence-to-structure mapping and an easy update procedure. The quality of the reconstructed sequences and the mapping were assessed by comparison with available MSD data.

As a front-end of our database, a Web interface was developed to annotate interactively UniProtKB/Swiss-Prot annotation from structures. Through this interface, an annotator can immediately know the structural properties of a residue in a given PDB structure (e.g. available coordinates, its association to an annotated site, a binding site or a protein interface). Another direct application of the mapping was a large-scale semi-automatic addition of easily computable structural features, such as disulfide bonds into UniProtKB/Swiss-Prot.

To summarize, the mapping presented here allows structural information to be used consistently for annotating UniProtKB/Swiss-Prot entries. It will be used both to suggest UniProtKB/Swiss-Prot annotation and to automatically check the consistency of existing annotation. The design of the mapping allows the analysis of UniProtKB/Swiss-Prot features in known structural contexts, which is an indispensable basis for further large-scale research studies.

**73 Henry van den Bedem<sup>1</sup>, Ankur Dhanik<sup>2</sup>, Jean-Claude Latombe, Ashley M. Deacon** <sup>1</sup>Joint Center for Structural Genomics / Stanford Synchrotron Rad. Lab., vdbedem@slac.stanford.edu <sup>2</sup>Department of Computer Science, Stanford U., ankurd@stanford.edu

### **Protein Dynamics from X-Ray Crystallography Data: Modeling Structural Heterogeneity with Inverse Kinematics**



**Short abstract:** We present a technique to model structural heterogeneity in protein main-chain fragments determined by X-ray crystallography. The technique also allows energetically plausible pathways between the conformations to be calculated, thus providing insight in the local dynamics of the main chain.

**Full abstract:** We present a method to model structural heterogeneity in protein main-chain fragments. A computer program using this method was developed to identify and simultaneously build multiple conformations of a fragment that fit the experimental X-ray data, together with their likelihoods. Additionally, the software can be used to calculate energetically plausible pathways between the

conformations, thus providing insight in the local dynamics of the main chain.

X-ray crystallography requires well-ordered protein crystals to produce diffraction images. Accurately determining the atomic coordinates of mobile fragments in a protein structure remains a challenge with this technique. For such fragments a simple temperature factor model, i.e. an averaged conformation together with a harmonic parameter to account for atomic vibrations, is often not adequate to explain the experimental data. For example, the electron density may indicate {it multimodal disorder}, where the protein main chain adopts two or more distinct conformations for a number of contiguous residues. Recognizing features in such areas of overlapping electron density is difficult, even for a trained crystallographer. There may also be regions where poor crystal packing leads to dynamic disorder of fragments in the macromolecule. In these regions the electron density is difficult or even impossible to interpret. At the same time, a protein's ability to bind to other molecules is typically facilitated by these mobile regions, and deformable fragments thus play a vital role in understanding the docking process and the protein's function. Such fragments of the protein main chain are better represented by an ensemble than a single, averaged conformation.

We have combined a fast inverse kinematics algorithm with real space, torsion angle refinement in a two stage approach to fit fragments to the electron density between two anchor points. The first stage samples a large number of closing conformations, guided by the electron density. These candidate conformations are ranked according to density fit, and then clustered. Top-ranking conformations from distinct clusters are simultaneously subjected to torsion angle, subchain refinement in the second stage. Optimization steps are projected onto the null space of the subchain, thus preserving rigid geometry and closure. Using conformational distance as a target function, this null space optimization yields pathways between conformations.

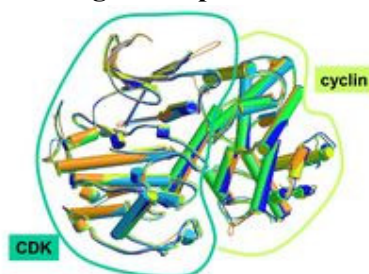
The algorithm was tested on synthetic data as well as experimental electron density. Using synthetic



atomic coordinates and calculated electron density, the software correctly identified and modeled two distinct conformers of length 8, separated by 4.4Å, to within 0.9Å and 0.6Å of the true conformers at a resolution of 1.5Å. Using real data, the software found two distinct conformers of length 8, separated by 2.6Å, to within 0.9Å and 0.5Å of the conformers deposited in the PDB (1VME) at a resolution of 1.8Å.

**74 Xueping Quan<sup>1</sup>, Peter Doerner<sup>2</sup>, Dietlind L Gerloff<sup>3</sup>** U. of Edinburgh, <sup>1</sup>x.quan@sms.ed.ac.uk  
<sup>2</sup>peter.doerner@ed.ac.uk <sup>3</sup>d.gerloff@ed.ac.uk

### **Partner Prediction for Transient Protein Interactions within sets of Paralogues: CDK-Cyclin homologue complexes in *A.thaliana***



**Short abstract:** To predict which specific CDK homologue interacts with which cyclin homologue in *Arabidopsis thaliana*, a comparative modelling strategy was applied to model the 3-D structures of these proteins. All-by-all docking of the model structures using the program ZDOCK, combined with additional selection criteria, yielded 19 most likely interacting CDK-cyclin pairs.

**Full abstract:** Introduction: CDKs (cyclin-dependent kinases) are Ser/Thr protein kinases that play an essential role in the regulation of eukaryotic cell proliferation, neuronal and thymus functions, and transcription in animals. Monomeric CDKs are inactive and require both association with a positive regulatory subunit, a cyclin, and phosphorylation of a conserved threonine residue that lies within the activation loop of the CDK, for full activity. The 3-D structures of CDKs and cyclins coming from other species, and/or other subfamilies, can be modelled using known and modelled structures of human CDKs and cyclins as templates. Previous studies and sequence sub-classification of the cyclin and CDK multigene families in the genome of the model plant *Arabidopsis thaliana* have revealed a minimum of 50 cyclin-like, and 35 CDK-like putative gene products. However, the sequence-structure relationships that determine the specificity of protein-protein interactions are not sufficiently understood at present to predict which specific CDK-cyclin pairings are likely to occur, and which are not. By focusing on this specific prediction challenge, we are working towards a better understanding of the biophysical principles governing protein-protein interactions in general, and towards developing computational methodology combining multiple components from several existing methods, that can be applied generally in this field.

**Methods:** ALL-BY-ALL DOCKING: Comparative models for 33 putative CDKs and 35 putative cyclins from *Arabidopsis thaliana* were produced with the program Modeller6, based on carefully adjusted multiple sequence alignments of each family [1]. The resulting structures were subjected to a large scale molecular docking experiment with ZDOCK [2] in which all CDK-cyclin combinations were considered.

**ADDITIONAL SELECTION CRITERIA:** Automated protein-protein docking results typically contain too many false positive complexes to be directly useful in practice. We therefore applied two additional criteria to select the best complexes from the ZDOCK result lists, based on:

- (i) Orientation - correct relative orientation of CDK and cyclin subunits in the complex;
- (ii) Interface – the electrostatic and hydrophobic properties at the interaction surface. Interface electrostatic correlation coefficients (ECC) and hydrophobic correlation coefficients (HCC) were calculated using the program MolSurfer [3].

**CALIBRATION:** For calibrating the interface criterion we used a positive, and a negative set of control data. The positive set consisted of 104 non-homologous, transient hetero-dimer complexes [4]. For the negative set 70 “non-complexes” were generated by using ZDOCK to combine proteins that do not normally interact (ZDOCK score > 60; interface size > 600 square angstroms). An optimal cutoff value for CEH (a combination of ECC and HCC) was chosen, based on the interface properties of these two sets.

**Results and Discussion:**

- Cross-validation, by randomly selecting 80% of data from each control set to derive the CEH cutoff, consistently yielded separation accuracies around 80% for the other 20% of control complexes. When we applied the entire modelling and interaction prediction approach to the well-characterized set of human CDKs and cyclins, 80% of the resulting predictions were in agreement with HRPD and Swiss-Prot annotation.
- All-by-all docking between 33 CDK models and 35 cyclin models yielded 83 pairs with outstanding ZDOCK scores ( $> 60$ ). Of these pairs, 59 are supported by correct orientation between the interacting subunits. Using CEH as an additional selection criterion, we retained 19 most likely interacting CDK-cyclin pairs in *Arabidopsis thaliana*. The most strongly predicted complex is formed between a close homologue of human CDK1/2/3, and a sequence most similar to human cyclinA (human CDK1/2/3-cyclinA are natural pairs). Another predicted complex has recently been confirmed experimentally.

#### References

1. Modeller: Sali & Blundell, 1993, *J.Mol.Biol.* **234**, 779-815
2. ZDOCK: Chen et al., 2003, *Proteins* **52**, 80-87
3. MolSurfer: Gabadoulle et al., 2003, *Nucl.Acids Res.* **31**, 3349-3351
4. Heterodimer Complexes: Ofra & Rost, 2003., *J.Mol.Biol.* **325**, 377-387

**75 Bruno Contreras-Moreira<sup>1</sup>, Julio Collado-Vides<sup>2</sup>** Centro de Ciencias Genómicas, UNAM, México,  
<sup>1</sup>contrera@ccg.unam.mx <sup>2</sup>collado@ccg.unam.mx

#### Comparative footprinting of DNA-binding proteins



**Short abstract:** We present the DNASITE software, designed for comparative footprinting of DNA-binding proteins based solely on structural data. While we find limitations to the approach, it can also provide useful predictions of binding sites.

**Full abstract:** Motivation: Comparative modelling is a computational method used to tackle a variety of problems in molecular biology and biotechnology. Traditionally it has been applied to model the structure of proteins on their own or bound to small ligands, although more recently it has also been used to model protein-protein interfaces. This

work is the first to systematically analyze whether comparative models of protein-DNA complexes could be built and be useful for predicting DNA binding sites.

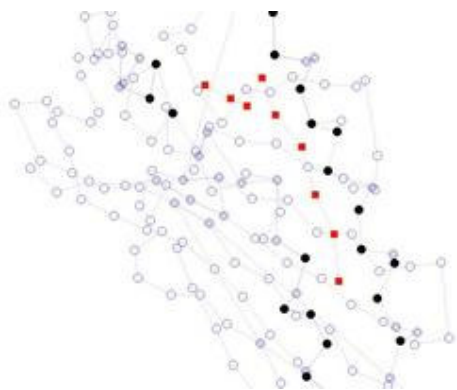
Results: First, we describe the structural and evolutionary conservation of protein-DNA interfaces, and the limits they impose on modelling accuracy. Second, we find that side-chains from contacting residues can be reasonably modeled and therefore used to identify contacting nucleotides. Third, the DNASITE protocol is implemented and different parameters are benchmarked on a set of 85 regulators from *Escherichia coli*. Results show that comparative footprinting can make useful predictions based solely on structural data, depending primarily on the interface identity with respect to the template used.

**78 Nebojsa Jojic<sup>1</sup>, Manuel J. Reyes-Gomez<sup>2</sup>, David Heckerman, Carl Kadie, Ora Schueler-Furman** Microsoft Research, <sup>1</sup>jojic@microsoft.com <sup>2</sup>manuelrg@microsoft.com

#### Learning MHC I - Peptide Binding

**Short abstract:** Motivated by the ability of a simple threading approach to predict MHC I - peptide binding, we developed a new improved structure-based model, which we call adaptive double threading. The parameters of the threading model are learnable, and both MHC and peptide sequences can be threaded onto structures of other alleles.

**Full abstract:** Motivated by the ability of a simple threading approach to predict MHC I - peptide binding, we developed a new and improved structure-based model for which parameters can be estimated from additional sources of data about MHC-peptide binding. In addition to the known 3D structures of a small number of MHC-peptide complexes that were used in the original threading approach, we included three other sources of information on peptide-MHC binding: (1) MHC class I sequences; (2) known binding energies for a large number of MHC-peptide complexes; and (3) an even larger binary dataset that contains information about strong binders (epitopes) and non-binders (peptides that have a low affinity



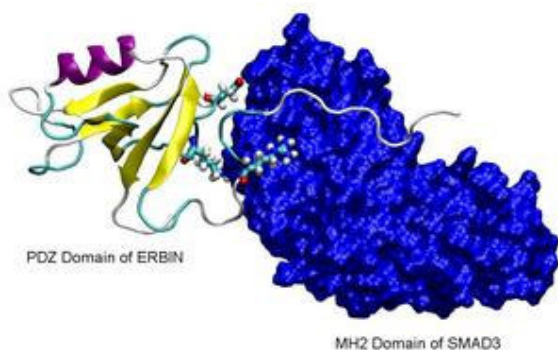
for a particular MHC molecule). Our model significantly outperforms the standard threading approach in binding energy prediction. In our approach, which we call adaptive double threading, the parameters of the threading model are learnable, and both MHC and peptide sequences can be threaded onto structures of other alleles. These two properties make our model appropriate for predicting binding for alleles for which very little data (if any) is available beyond just their sequence, including prediction for alleles for which 3D structures are not available. The ability of our model to generalize beyond the MHC types for which training data is available also separates our approach from epitope prediction methods which treat MHC alleles as symbolic types, rather than biological sequences. We used the trained binding energy predictor to study viral infections in 246 HIV patients from the West Australian cohort, and over 1000 sequences in HIV clade B

from Los Alamos National Laboratory database, capturing the course of HIV evolution over the last 20 years. Finally, we illustrate short-, medium-, and long-term adaptation of HIV to the human immune system.

**79 Chavent Matthieu<sup>1</sup>, Claire Nourry<sup>2</sup>, Nadine Deliot, Jean P. Borg, Bernard Maigret**

<sup>1</sup>Henri Poincaré U., UMR 7565, eDAM team, matthieu.chavent@edam.uhp-nancy.fr <sup>2</sup>U119 INSERM and Institut Paoli-Calmettes, Laboratory of Molecular Pharmacology, nourry@marseille.inserm.fr

### **Use of Docking Tools and Molecular Dynamic Simulations to Predict the Interaction Between the MH2 domain of SMAD3 and the PDZ Domain of ERBIN**



**Short abstract:** Erbin has been identified as a binding partner for Smad3, a major component of the transforming growth factor-beta signaling pathway. We present how the use of bioinformatics and molecular modelling tools such as molecular dynamics helped us to conceptualize the association between the Smad3 MH2 and Erbin PDZ domains.

**Full abstract:** Erbin was originally identified as a protein that interacts with the receptor tyrosine kinase ErbB2 (also known as HER-2 or Neu) and plays a role in its signalling activity in epithelial and neuronal cells. This protein contains 16 leucine-rich repeats in its amino terminus and a PDZ domain in its carboxy terminus. PDZ domains are widely distributed protein-protein interaction domains in the human genome. In humans, about 540 PDZ domains are present in proteins that

have signalling, scaffolding and regulatory functions. Recently, Erbin has been identified as a binding partner for Smad3, a major component of the transforming growth factor-beta (TGF beta) signaling pathway.

The Smad proteins constitute a family of intracellular proteins that function as signal transducers for the TGF beta superfamily. The eight human Smads, Smad1 to Smad8, can be grouped into three subfamilies: the receptor-regulated Smads (R-Smads), the common Smads (Co-Smads) and the inhibitory Smads (I-Smads). Smad3 shares a common domain organization with other R-smads and Smad 4, consisting in an amino-terminal DNA-binding domain (MH1 domain) and a carboxy-terminal effector domain (MH2 domain) separated by a linker region. The MH2 domain is a versatile protein-protein interaction module interacting with a long list of partners.

It has been discovered that Erbin and Smad3 form a protein complex in vivo and directly interact in vitro. However it remains to determine the modalities of interaction between Erbin and Smad3 at the molecular level. Biochemical experiments demonstrate that the MH2 domain of Smad3 is required for this interaction, and that the carboxyterminal region of Erbin encompassing residues 914 to 1257, plus the PDZ domain are involved in this interaction. Furthermore, a mutagenic analysis has shown that the classical binding pocket of the PDZ domain plays no role in the interaction. In this poster, we present how the use of the bioinformatic and molecular modelling tools (alignments and structure analysis programs) helped us to conceptualize the association between the Smad3 MH2 and the Erbin PDZ domains and to propose a model for the association of the two partners. Additionally, we used the ClusPro automatic docking web server to create another model association that agrees with the few available experimental datas. We then carried out a molecular dynamic of 20 ns for the two models of association to evaluate their stability. These molecular dynamics calculations show that the two models converge to a unique position in which we defined key residues in both Smad3 and Erbin important for the interaction. We are currently in the process to mutate these residues to validate our proposal.

**80 Alexandre Beautrait<sup>1</sup>, Bernard Maigret<sup>2</sup>** eDAM group, UMR CNRS/UHP 7565, <sup>1</sup>alexandre.beautrait@edam.uhp-nancy.fr <sup>2</sup>bernard.maigret@edam.uhp-nancy.fr

### **VSM-G: the Virtual Screening Manager Platform for Computational Grids. Use for the Identification of Putative Ligands of the Liver X Receptor.**



**Short abstract:** The VSM-G platform performs high throughput virtual screening based on ligand- and receptor-based algorithms running on computational grids. A central molecular funnel composed of successive filters allows the screening of large molecular libraries for compounds prioritization in experimental assays. Illustrations deal with the LXR target protein involved in cholesterol regulation.

**Full abstract:** The search for novel therapeutics is time-consuming and expensive, so that any method able to speed up the process is welcome. In the last several years, virtual screening techniques have evolved from a side-show curiosity to a central partner in many drug development strategies. Virtual screening provides numerous advantages, including the speed with which one can screen a large molecular library and a relatively small capital investment to get started compared to the cost of an in vitro HTS program. Presently, virtual screening methods are proving to be a good complement to high-

throughput screening experiments, by limiting the number of molecules to be tested and therefore the cost of screening campaigns aimed at identifying putative lead compounds.

In such a context, we have developed a Java-based standalone platform, called VSM-G. Its main goal is the prioritization of compounds for experimental testing, by performing high throughput virtual screening. To satisfy this task as efficiently as possible, the software integrates a chain of tools combining both ligand- and receptor-based strategies. The former is implemented as modules such as substructure search to pre-screen molecular databases in order to restraint the number of compounds to consider subsequently.

This optional operation may precede the central element of the platform, a multi-step screening funnel which presents several layers of filtering. This approach consists in the succession of different selection methods used sequentially for evaluating the possible affinity of each compound with the protein targets. This stepwise process starts with first geometrical considerations for ligand-cavity matching followed by more efficient flexible docking algorithms and ending with molecular dynamics simulations.

At each step a fair percentage of non appropriate molecules may therefore be eliminated, so that the system is able to handle several million compounds, eventually yielding to a small set of putative leads. The capability of screening several million compounds against several hundred biological targets, within a reasonable time, rests on a grid module that spreads the calculations over computational grids composed of PC clusters.

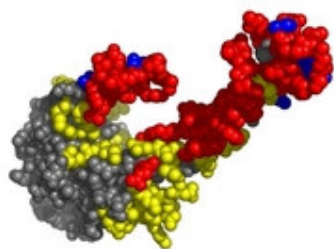
Illustrations will describe the basic elements of the platform and also showcase its capabilities through an example of biological interest concerning a particular nuclear receptor, namely the Liver X Receptor LXR  $\alpha$ . This case illustrates a general effort towards the discovery of new pharmaceuticals for treating cardiovascular diseases by targeting this receptor involved in cholesterol regulation.

**81 Michael Terribilini<sup>1</sup>, Jeffrey Sander<sup>2</sup>, Byron Olson<sup>3</sup>, Jae-Hyung Lee, Robert Jernigan, Vasant Honavar, Drena Dobbs**

Iowa State University, <sup>1</sup>terrible@iastate.edu <sup>2</sup>jdsander@iastate.edu <sup>3</sup>byron.olson@gmail.com

## A Computational Method to Identify Amino Acid Residues Involved in Protein-RNA

### Interactions



**Short abstract:** Protein-RNA interactions are vital for many biological processes. To the best of our knowledge, this is the first reported attempt to exploit both amino acid sequence and three-dimensional structure for prediction of RNA-binding residues in proteins.

**Full abstract:** Protein-RNA interactions are vitally important in a wide range of biological processes, including regulation of gene expression, protein synthesis, and replication and assembly of many viruses. The ability to reliably predict which residues of a protein directly contribute to RNA binding would significantly enhance our understanding of how proteins recognize RNA and potentially generate new strategies for

clinical intervention in both genetic and infectious diseases.

We have developed a machine learning approach for predicting which amino acids of an RNA-binding protein are involved in protein-RNA interactions, using either amino acid sequence only or a combination of both sequence and structure-derived information as input. Interfaces from known protein-RNA complexes in the PDB were extracted to generate a non-redundant set of 147 RNA-binding protein chains. Using this dataset, several types of classifiers were trained to predict which residues in a given RNA binding protein are located at the protein-RNA interface. The first classifier



uses only the amino acid sequence as input, whereas the second classifier uses structural information as input. We then combine the results of these two individual classifiers to obtain the final prediction. The combined classifier identifies interface residues with 87% overall accuracy, correlation coefficient of 0.37, specificity for interface residues of 57% and sensitivity for interface residues of 33%. An example is shown in Fig. 1. The classifier can be calibrated to increase the specificity or sensitivity of interface residue prediction for specific functional classes of RNA-binding proteins. To the best of our knowledge, this is the first reported attempt to exploit both amino acid sequence and three-dimensional structure for prediction of RNA-binding residues in proteins. We have applied the sequence-based classifier to several clinically important proteins for which experimental structure information has been lacking, e.g. HIV-1 Rev and the human telomerase reverse transcriptase (hTERT) protein. In both cases, the predicted RNA-binding residues agree with the biochemically mapped RNA-binding regions.

## Registrant list:

<b>fname</b>	<b>lname</b>	<b>org</b>	<b>city</b>	<b>state</b>	<b>country</b>	<b>email</b>
Iris	Antes	Max-Planck-Institute	Saarbruecken		GERMANY	antes@mpi-sb.mpg.de
Daniel	Barker	U. of St Andrews	Fife		UK	db60@st-andrews.ac.uk
Alexandre	Beautrait	UHP/CNRS	Nancy		FRANCE	alexandre.beautrait@edam.uhp-nancy.fr
Andreas	Bernsel	Stockholm Bioinfo Center	Stockholm		SWEDEN	andreas.bernsel@sbc.su.se
Talapady	Bhat	NIST	Gaithersburg	MD	USA	bhat@nist.gov
Richard	Bickerton	U. of Cambridge	Cambridge		UK	richard@cryst.bioc.cam.ac.uk
Karsten M.	Borgwardt	Ludwig-Maximilians-U.	Munich	Bavaria	GERMANY	kb@dbis.ifi.lmu.de
Philip	Bourne	UCSD	La Jolla	CA	USA	pbourne@ucsd.edu
James	Bradford	U. of Leeds	Leeds	W Yorkshire	UK	bmbjrb@bmb.leeds.ac.uk
Steven	Brenner	U. of California	Berkeley	CA	USA	brenner@compbio.berkeley.edu
Janusz	Bujnicki	International Institute	Warsaw		POLAND	iamb@genesilico.pl
Andrew	Bulpitt	U. of Leeds	Leeds	West Yorkshire	UK	a.j.bulpitt@leeds.ac.uk
Anke	Busch	U. Freiburg	Freiburg	-	GERMANY	anke.busch@informatik.uni-freiburg.de
Dirksen	Bussiere	Chiron Corporation	Emeryville	CA	USA	dirksen_bussiere@chiron.com
Jonas	Carlsson	IFM Bioinfo	Linkoeeping		SWEDEN	jonca@ifm.liu.se
Matthieu	Chavent	UHP/CNRS	Vandoeuvre Les Nancy		FRANCE	matthieu.chavent@edam.uhp-nancy.fr
Leo	Cheung	Loyola U. Medical Center	Maywood	IL	USA	lcheung@lumc.edu
Wah	Chiu	Baylor College of Med	Houston	TX	USA	wah@bcm.edu
Bruno	Contreras- Moreira	UNAM	Cuernavaca	Morelos	MEXICO	contrera@ccg.unam.mx
Fabrice	David	Swiss Institute of Bioinfo	Geneva		SWITZERLA ND	fabrice.david@isb-sib.ch
Charlotte	Deane	Oxford U.	Oxford	Oxfordshire	UK	deane@stats.ox.ac.uk
Antonio	Del Sol	Fujirebio Inc	Tokyo		JAPAN	ao-mesa@fujirebio.co.jp
Frank	DiMaio	U. of Wisconsin	Madison	WI	USA	fpdimaio@wisc.edu
Sara	Dobbins	Imperial College London	London		UK	sara.dobbins@imperial.ac.uk
Zsuzsanna	Dosztanyi	Institute of Enzymology	Budapest		HUNGARY	zsuzsa@enzim.hu
Sumeet	Dua	Louisiana Tech U.	Ruston	LA	USA	sdua@coes.latech.edu
Diana	Ekman	Stockholm Bioinfo Center	Stockholm		SWEDEN	diaek@sbc.su.se
Sebastian	Fernandez- Alberti	Universidad Nacional	Bernal	Buenos Aires	ARGENTINA	seba@unq.edu.ar
Andras	Fiser	Albert Einstein College	Bronx	NY	USA	andras@fiserlab.org
Robson	Francisco de Souza	U. de Sao Paulo	Sao Paulo		BRAZIL	robfsouza@gmail.com
Pablo	Freire	LNCC - National Laboratory for	Petropolis	Rio de Janeiro	BRAZIL	pablorf@lncc.br
Thomas	Funkhouser	Princeton U.	Princeton	NJ	USA	funk@cs.princeton.edu
Oxana	Galzitskaya	Institute of Protein Research	Moscow		RUSSIA	ogalzit@vega.protres.ru
Dietlind	Gerloff	U. of Edinburgh	Edinburgh	Scotland	UK	Dietlind.Gerloff@ed.ac.uk
Nick	Grishin	HHMI/UT Southwestern	Dallas	TX	USA	grishin@chop.swmed.edu
Jorge	H Fernandez	LabInfo LNCC MCT	Petropolis	RJ	BRAZIL	jorgehf@lncc.br
Vasant	Honavar	Iowa State U.	Ames	IA	USA	honavar@cs.iastate.edu
Bingding	Huang	Biotech, TU Dresden	Dresden		GERMANY	bhuang@biotech.tu-dresden.de
Fabien	Huard	Macquarie U.	North Ryde	NSW	AUSTRALIA	fhuard@efs.mq.edu.au
Muyideen						
A.	Imam	StStephen Institute of	Banjul	KN	GAMBIA	stephen_institute@yahoo.co.uk
Benjamin	Jefferys	Imperial College	London	London	UK	benjamin.jefferys@imperial.ac.uk
Chan-seok	Jeong	Korea Advanced Institute of Science and Technology	Yuseong-gu	Daejeon	S. KOREA	blna999@kaist.ac.kr
Anna	Johansson	Stockholm Bioinfo Center	Stockholm		SWEDEN	annj@sbc.su.se
Nebojsa	Jojic	Microsoft Research	Redmond	WA	USA	jojic@microsoft.com

Inge	Jonassen	U. of Bergen	Bergen		NORWAY	inge@ii.uib.no
Kevin	Karplus	U. of California	Santa Cruz	CA	USA	karplus@soe.ucsc.edu
Firas	Khatib	U. of California	Santa Cruz	CA	USA	fkhatib@ucsc.edu
	Krishnamoort					
Bala	hy	Washington State Univ	Pullman	WA	USA	kbala@wsu.edu
Paula	Kuser-Falceo	Embrapa Informatica Agrop	Campinas	Sao Paulo	BRAZIL	paula@cnptia.embrapa.br
Braddon	Lance	Macquarie U.	Croydon	NSW	AUSTRALIA	blance@efs.mq.edu.au
Melissa	Landon	Boston U.	Boston	MA	USA	mlandon@bu.edu
Minho	Lee	Korea Advanced Institute	Yuseong-gu	Daejeon	S. KOREA	MinhoLee@kaist.edu
Thomas	Lengauer	Max-Planck Institute	Saarbrücken	.	GERMANY	lengauer@mpi-sb.mpg.de
Stefano	Lise	U. College London	London		UK	lise@biochem.ucl.ac.uk
			Vandoeuvre les			
Bernard	Maigret	H. Poincare U. - Nancy	NANCY Cede	Lorraine	FRANCE	bernard.maigret@edam.uhp-nancy.fr
Kimmo	Mattila	CSC - Scientific Computing Ltd.	Espoo	Uusimaa	FINLAND	kimmo.mattila@csc.fi
Francisco	Melo	Universidad Catolica	Santiago	RM	CHILE	fmelo@bio.puc.cl
John	Moult	CARB, U. Maryland		MD	USA	jmoult@tunc.org
Rafael	Najmanovich	EMBL-EBI	Cambridge		UK	rafael.najmanovich@ebi.ac.uk
Chris	Needham	School of Computing	Leeds	W Yorkshire	UK	chrisn@comp.leeds.ac.uk
Gustavo	Parisi	Universidad Nacional Quilmes	Bernal	Buenos Aires	ARGENTINA	gustavo@unq.edu.ar
Sung Hee	Park	ETRI	Daejeon		S. KOREA	sunghee@etri.re.kr
Marcin	Pawlowski	International Institute	Warsaw		POLAND	marcinp@genesilico.pl
Riccardo	Percudani	Università di Parma	Parma	na	ITALY	riccardo.percudani@unipr.it
Marco	Punta	Columbia U.	New York	NY	USA	mp2215@columbia.edu
Stephane	Redon	INRIA	Saint-Ismier	Rhone-Alpes	FRANCE	stephane.redon@inria.fr
Manuel	Reyes-Gomez	Microsoft Research	Redmond	WA	USA	manuelrg@microsoft.com
Walter	Rocchia	NEST	Pisa		ITALY	w.rocchia@sns.it
Peter	Roegen	Tech. Univ. Denmark	Kgs. Lyngby		DENMARK	Peter.Roegen@mat.dtu.dk
Torbjorn	Rognes	U. of Oslo	Oslo		NORWAY	torbjorn.rognes@medisin.uio.no
Ingo	Ruczinski	Johns Hopkins U.	Baltimore	MD	USA	ingo@jhu.edu
John	Rux	Wistar Institute	Philadelphia	PA	USA	rux@wistar.upenn.edu
Gisle	Saelensminde	U. of Bergen	Bergen		NORWAY	gisle@cbu.uib.no
Ilan	Samish	U. Pennsylvania	Philadelphia	PA	USA	samish@sas.upenn.edu
Oliver	Sander	Max-Planck-Institute	Saarbrücken	Saarland	GERMANY	osander@mpi-sb.mpg.de
Jeffery	Saven	U. Pennsylvania	Philadelphia	PA	USA	saven@sas.upenn.edu
Avner	Schlessinger	Columbia U.	New York	NY	USA	as2067@columbia.edu
Masakazu	Sekijima	AIST	Tokyo		JAPAN	m.sekijima@aist.go.jp
Marian	Seto	Berlex Biosciences	Richmond	CA	USA	marian_seto@berlex.com
Sven	Siebert	Albrecht-Ludwigs Univ	Freiburg		GERMANY	siebert@informatik.uni-freiburg.de
Christopher	Southan	AstraZeneca R&D	Mølndal	Göteborg	SWEDEN	Christopher.southan@astrazeneca.com
Kristen	Taylor	Joint Genome Institute	Walnut Creek	CA	USA	kmtaylor@lbl.gov
Janet	Thornton	EMBL-EBI	Cambridge	Cambs	UK	helen@ebi.ac.uk
Nicolas	Tsesmetzis	John Innes Centre	Norwich	Norfolk	UK	nicolas.tsesmetzis@bbsrc.ac.uk
Alfonso	Valencia	Spanish Natl Cancer Res, CNIO	Madrid		SPAIN	avalencia@cnio.es
	Van den					
Henry	Bedem	Stanford Linear Accelerator Center	Menlo Park	CA	USA	vdbedem@slac.stanford.edu
Stella	Veretnik	UCSD	La Jolla	CA	USA	veretnik@sdsc.edu
Håkan	Viklund	Dept of Biochemistry	Stockholm		SWEDEN	hakany@sbcsu.se
				Baden-		
Sebastian	Will	U. of Freiburg	Freiburg	Württemberg	GERMANY	will@informatik.uni-freiburg.de
Scooter	Willis	U. of Florida	Gainesville	FL	USA	willishf@ufl.edu
Graham	Wood	Macquarie U.	Sydney	NSW	AUSTRALIA	gwood@efs.mq.edu.au
Monika	Wood	Promega Corporation	Madison	WI	USA	mwood@promega.com