3DSIG 2009: The 5th Structural Bioinformatics and Computational Biophysics

An ISMB satellite meeting

June 27-28, 2009
Stockholm - Sweden

http://www.ebi.ac.uk/~rafi/3dsig09

Scientific Committee:

John Moult (chair), Michael Levitt, Gunnar von Heijne, Nir ben Tal, Arne Elofsson, David Jones, Haim Wolfson

Organising Committee:

Rafael Najmanovich & Ilan Samish
International Society for Computational biology

The European Molecular Biology Organization

The Human Frontiers Science Program Journal

Molecular Graphics Modelling Society

Roche

Wiley

World Scientific

CRC Press
## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>4</td>
</tr>
<tr>
<td>Keynote presentation Abstracts</td>
<td>6</td>
</tr>
<tr>
<td>Illustrated abstract table of contents</td>
<td>12</td>
</tr>
<tr>
<td>Abstracts</td>
<td>17</td>
</tr>
<tr>
<td>List of participants</td>
<td>88</td>
</tr>
<tr>
<td>Time</td>
<td>ID</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
</tr>
<tr>
<td>8:55</td>
<td></td>
</tr>
<tr>
<td>9:00</td>
<td>K01</td>
</tr>
<tr>
<td>9:30</td>
<td>11</td>
</tr>
<tr>
<td>9:50</td>
<td>20</td>
</tr>
<tr>
<td>10:10</td>
<td>51</td>
</tr>
<tr>
<td>10:30</td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td>K02</td>
</tr>
<tr>
<td>11:30</td>
<td>29</td>
</tr>
<tr>
<td>11:50</td>
<td>50</td>
</tr>
<tr>
<td>12:10</td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>K03</td>
</tr>
<tr>
<td>15:30</td>
<td>26</td>
</tr>
<tr>
<td>15:50</td>
<td>27</td>
</tr>
<tr>
<td>16:10</td>
<td>31</td>
</tr>
<tr>
<td>16:30</td>
<td></td>
</tr>
<tr>
<td>17:00</td>
<td>8</td>
</tr>
<tr>
<td>17:20</td>
<td>12</td>
</tr>
<tr>
<td>17:40</td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td></td>
</tr>
<tr>
<td>18:45</td>
<td></td>
</tr>
<tr>
<td>19:30</td>
<td></td>
</tr>
<tr>
<td>21:00</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>ID</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>9:00</td>
<td>K04</td>
</tr>
<tr>
<td>9:30</td>
<td>17</td>
</tr>
<tr>
<td>9:50</td>
<td>35</td>
</tr>
<tr>
<td>10:10</td>
<td>53</td>
</tr>
<tr>
<td>10:30</td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td>K05</td>
</tr>
<tr>
<td>11:30</td>
<td>15</td>
</tr>
<tr>
<td>11:50</td>
<td>32</td>
</tr>
<tr>
<td>12:10</td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>K06</td>
</tr>
<tr>
<td>15:30</td>
<td>39</td>
</tr>
<tr>
<td>15:50</td>
<td>43</td>
</tr>
<tr>
<td>16:10</td>
<td>61</td>
</tr>
<tr>
<td>16:30</td>
<td></td>
</tr>
<tr>
<td>17:00</td>
<td>37</td>
</tr>
<tr>
<td>17:20</td>
<td>54</td>
</tr>
<tr>
<td>17:40</td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td></td>
</tr>
<tr>
<td>18:35</td>
<td></td>
</tr>
</tbody>
</table>
KEYNOTE PRESENTATION ABSTRACTS

K01. The Protein Universe: A Current Snapshot

Michael Levitt, Stanford University, USA.

All currently known sequences in the Protein Universe are analyzed in terms of families that have single or multi-domain architectures, and whether they have a known three-dimensional structure. Growth of new single domain families is very slow: almost all growth comes from new multi-domain architectures that are combinations of domains recognized by about 15,000 sequence profiles.

The protein universe is the set of all proteins of all organisms. Here all currently known sequences are analyzed in terms of families that have single or multi-domain architectures, and whether they have a known three-dimensional structure. Growth of new single domain families is very slow: almost all growth comes from new multi-domain architectures that are combinations of domains recognized by about 15,000 sequence profiles. Single-domain families are mostly shared by the major groups of organisms whereas multi-domain architectures are specific and account for species diversity. There are known structures for a quarter of the single domain families and over 70% of all sequences can be partially modeled thanks to their membership of these families.

Figure: Cartoons of sequence space in which area is proportional to the number of sequences or sequence families in that region. Sequences not characterized by any merged CDART sequence profile are the dark matter of the protein universe (23% of 7,300,000, the grey core). (a) The Repetitious Sequence Universe contains all sequences. Most characterized sequences (88%, orange area) have single domain architectures (SDA), where one region of the sequence is matched by a sequence profile (colored bar on black line). The remainder (12%, blue area) have multi-domain architectures (MDA) with more than one region of the sequence is matched (several colored bars on sequence). Over three-quarters (76%) of the SDA sequences are matched by a sequence profile family that has a known three-dimensional structure and 4% of the SDA sequences were solved by structural genomics (brown area, cross-hatching indicates domain of known structure). (b) The Unique Sequence Universe contains all sequence families. 86% of the families are MDA and other 14% are SDA. 32% of SDA sequence families have a known structure, with a fifth of these from structural genomics. For 49% of the MDAs, all domains have a known structure (cross-hatched) and another 42% have at least one domain with a known structure (part PDB). (c) shows the numbers of sequences in the corresponding regions of (a), whereas (d) shows the number of families in the corresponding regions of (b). This copywrite of this figure belongs to the Proceedings of the National Academy of Sciences, USA.

K02. Inferring function from disorder, domain and transmembrane predictions

Anna E. Lobley, Tim Nugent and David T. Jones, Bioinformatics Group, Dept. of Computer Science, University College London, U.K.

The impact of high throughput sequencing technologies since the 1980’s has produced over 100 billion base pairs of DNA sequence, cataloguing the genetic material of more than 800 organisms. Genome sequences provide information not only for a complete set of genes and their precise locations in the chromosome, but also help to define the core proteome i.e. the set of functional proteins that are the work horse components of living cells. In this post-sequencing era, one of the key challenges is to predict the function of a protein given its amino acid sequence. Most automated function prediction methods rely upon identifying well annotated sequence and structural homologues to transfer annotations to uncharacterized proteins. Sequence similarity based methods are relatively successful at annotating homologous proteins, however, they are not applicable to annotating orphan proteins or proteins whose relatives are not themselves functionally annotated. Currently, around 35% of proteins cannot be accurately annotated by homology-based transfer methods, which highlights the need for function prediction methods that are independent of sequence similarity.
The first approach I will discuss, FFPred, adopts a machine learning classification approach to perform function prediction in protein feature space using characteristics predicted from amino acid sequence. The features are scanned against a library of Support Vector Machines representing over 300 Gene Ontology classes and probabilistic confidence scores returned for each annotation term. The GO term library has been modelled on human protein annotations, however benchmark performance testing showed robust performance across higher eukaryotes. FFPred offers important advantages over traditional function prediction servers in its ability to annotate distant homologues and orphan protein sequences, and achieves greater coverage and classification accuracy than other feature based prediction servers particularly due to the incorporation of patterns of native disorder in the target protein. Such natively unstructured regions are a common feature of eukaryotic proteomes, particular those of higher organisms where between 30-60% of proteins are predicted to contain long stretches of disordered residues. Pattern analysis of the distributions of lengths and positions of these disordered regions demonstrated that the functions of intrinsically disordered proteins are indeed dependent on these parameters.

In this talk I will also be describing more recent work we have been doing to predict protein function by combining genome-wide predictions of protein disorder, transmembrane topology, domain patterns with other sources of biological information such as microarray data and protein-protein interaction features (Figure 1). To better achieve these goals we have developed improved versions of two of our widely used prediction tools: MEMSAT-SVM (transmembrane topology prediction) and pDomTHREADER (domain-based fold recognition), and these recent improvements will also be briefly outlined.

References


Membrane proteins present both experimentalists and computational scientists with special challenges. For the biochemist, they are often difficult to overexpress, purify and study in isolation. These difficulties are glaringly obvious in protein crystallography, with only a few percent of all known crystal structures representing membrane proteins. Yet, for the computational biologist the field of membrane protein structure/function prediction is fertile ground, partly because the experimental studies are so difficult, partly because the structural characteristics of membrane proteins are in some respects simpler than those of soluble proteins.

From the computational point of view, the main areas of study are topology prediction, homology modeling, 3D structure prediction constrained by ‘low-resolution’ experimental data, and molecular dynamics analysis of lipid-embedded proteins. Up until recently, almost all of the prediction methods have been based on statistical analysis of known structures and more or less sophisticated machine-learning approaches. We have taken a different approach, trying to derive basic free energies of membrane insertion for the different amino acids from in vivo studies of membrane insertion of transmembrane α-helices, (Fig. 1). Using this kind of experimental data, we have been able to develop a very simple topology prediction scheme with only two free parameters that
performs on par with the best HMM-based methods. Our results further show that the ‘hydrophobicity threshold’ required for efficient membrane insertion of transmembrane helices in multi-spanning membrane proteins can be considerably higher than that seen for single-spanning membrane proteins, strongly suggesting that tertiary interactions between transmembrane helices can be important for membrane insertion.

References


K04. The evolutionary history of domains and multi-domain proteins.

Arne Elofsson, Diana Ekman, Åsa Björklund and Rauan Sagit. Stockholm University, Sweden.

More than 70% of the proteins in the human proteome consist of multiple domains and about 17% of the proteins contain repeated domains. We have found that the evolution of most multi-domain proteins can be explained by simple addition of single domains at the N or C-terminus. However, one important exception to this rule exists, the evolution of repeating domains. In our studies we have found that the evolution of most multi-domain proteins can be explained by simple addition of single domains at the N or C-terminus, while the addition of multiple domains simultaneously or central domains seems to be rare. However, one important exception to this rule exists, the evolution of repeating domains. Here it seems to be the rule that multiple domains are added simultaneously, see Figure 2. Further, their evolution works through internal duplications. The evolution of repeat containing proteins has played a key role in the development of complex regulatory systems and signaling required for multicellularity.

What is the molecular mechanism that underlies the evolution of repeating proteins? We have observed that often in a repeating protein two neighboring domains are less similar to each other than domains further apart 1. We believe that this is the cause of a molecular mechanism that copies several domains at a time. However, Clarke and co-workers have proposed that the reason for this is that neighboring evolve to be different to avoid miss-folding and aggregation4. The basic controversy rules down to if the observed pattern is due to a selective pressure at the protein domain level or if it is a result of a specific copying mechanism on the DNA level. We believe that this question can be answered by a detailed study of repeating proteins, in particular the Nebulin family.

Secondly, we will discuss differences in domain architecture evolution in different species. Most eukaryotic proteins consist of multiple domains, created through gene fusions or internal duplications. However, the mechanisms underlying the evolution of multidomain proteins are not very well studied. We have studied the evolution of multidomain architectures, guided by evolutionary information in the form of a phylogenetic tree. Our results show that Pfam domain families and multidomain architectures have been created with comparable rates (0.1-1 per My). The major changes in domain architecture evolution have occurred in the process of multicellularization and within the metazoan lineage. In contrast, creation of domains seems to have been frequent already in the early evolution. Further, most of the architectures have been created from older domains or architectures, whereas novel domains are mainly found in single domain proteins. However, a particular group of exon bordering domains may have contributed to the rapid evolution of novel multidomain proteins in metazoan organisms. Finally, multidomain architectures have evolved predominantly through insertions of domains, whereas domain deletions are less common.

The domain concept is a fundamental concept in our understanding of proteins. It is assumed that a domain is “an independently folding unit” as well as “evolutionary independent”. More than 70% of the proteins in the human proteome consist of multiple domains and about 17% of the proteins contain repeated domains. In the human cells repeated domains are essential for signaling and regulation and are frequent among protein interaction hubs. Here, we will discuss mechanisms involved in the evolution of domains and domain architectures. Also, for large regions of many proteins, and even entire proteins, no homology to known domains or to other proteins can be detected. These sequences are often referred to as orphans. However, it is believed that de novo creation of coding sequences is rare in comparison to mechanisms such as domain shuffling and gene duplication, hence most sequences should have homologs in other genomes. Surprisingly, the large amount of orphans is sustained in spite of a rapid increase of available genomic sequences.
Finally, we will discuss the creation of novel domains, particularly in yeast. Here, we used two approaches for detection of orphan domains and orphan proteins, which gave similar estimates of the rate of protein innovation in S. cerevisiae. They agreed that about 3-6% of the residues belong to sequences specific to the Saccharomyces lineage, whereas less than 2% are species specific. At the S. cerevisiae level the orphan proteins outnumber the orphan domains, and only about twenty orphan domains were predicted. However, the opposite was observed at larger evolutionary distances, which might to some extent be a result of alignment problems. We found that to detect truly de novo created orphans it is necessary to compare the proteins from very closely related species. Most of the detected orphan domains exist in one single copy in the S. cerevisiae genome, and are preferentially short sequences located at the protein termini. Their terminal location suggests a possible mechanism for their creation, involving changes of start and stop codons. In addition, the Saccharomyces specific orphan domains contain a high amount of low complexity regions and disorder, which suggests that nucleotide repetitions could be involved in de novo creation. However, increased evolutionary rates are associated with disorder, and as the amount of disorder is low in the S. cerevisiae orphan domains, disordered sequences might not be orphans, but just be more difficult to align. In contrast, orphan proteins appear to be structurally similar to older proteins, although shorter. In particular many of the species specific proteins are extremely short, and their functions are largely uncharacterized.

We also noted that two main mechanisms were responsible for de novo creation of coding sequences. Frequently, one protein terminus is altered by introduction of novel start or stop sites. In other cases the length difference between proteins is strongly correlated with an increase in the amount of disorder in the protein. Hence, changes within disordered regions seems to be a major factor in length increase/decrease. Finally, tandem repeats of short nucleotide sequences were observed in a number of increasing proteins, which suggests that this is an important mechanism in protein elongation. In contrast to changes in start and stop codons, this has also expanded proteins at internal positions.

References

Figure 1: Pattern of internal domain duplications in the chicken protein, ENSGALP0000020382, with 70 repeating nebulin domains (Pfam). The intensity of the squares is related to alignment scores with darker intensity indicating higher scores and the numbers on both axes indicate the domains in N-to-C terminal orientation. The protein has strong similarity between every seventh domain, indicating that a cassette of seven domains has been duplicated several times.

K05. Computational Methods for the Modeling of Large Macromolecular Assemblies

Haim J. Wolfson, School of Computer Science, Tel Aviv University, Israel.

Modelling of multimolecular assemblies is crucial for the understanding of cellular function. Nevertheless, most of the structures in the PDB are either monomers or dimers. The yeast cell, for example, contains approximately 800 distinct core complexes of 5 proteins, on the average, most of which have not yet been structurally characterized. In parallel, the various Structural Genomics efforts and the improvement in homology modeling techniques provide a wealth of single protein atomic resolution structures. In addition, recent developments in experimental techniques, such as cryo-EM, FRET, SAXS provide low resolution information on large macromolecular complexes and distance constraints between the interacting units. Development of efficient algorithms, which integrate the high and low resolution data, in order to model macromolecular complexes at atomic resolution is a key Structural Bioinformatics task.

We present several methods, which tackle the Multimolecular Assembly task by integrating experimental protein data at different resolutions. Among these methods are:

1. **CombDock** – a method for combinatorial docking of atomic resolution structures (or high quality models) to predict their assembly. A version which exploits distance constraints will be discussed.
2. **EMatch** – accepts as input an intermediate resolution cryo-EM density map of the complex and fits PDB protein fold representative structures into this map to detect the assembly proteins and their pose.
3. **MultiFit** – accepts as input a low resolution (~20Å) density map of the assembly and atomic resolution structures or models of the participating proteins and assembles the “puzzle” by a combination of fitting and docking techniques.

All the algorithms are run-time efficient, have been implemented, and performed well on available benchmark data.

References:
K06. Modeling and validation of transmembrane protein structures (Based on a chapter written for the book "Protein Structure Prediction" Edited by Huzefa Rangwala and George Karypis)

Maya Schushan and Nir Ben-Tal, Tel Aviv University, Israel.

In comparison to water-soluble proteins, it is easier to predict the 3D-structure of helical transmembrane proteins because of their limited topologies. However, the usefulness of standard homology modeling techniques is limited because the library of relevant templates is rather small. Evolutionary information is useful both for structural modeling and validation.

Transmembrane (TM) proteins comprise some 15% to 30% of the proteome, and the number of reported structures has grown rapidly over the past decade. Nevertheless, owing to technical difficulties, relatively few of these structures have been determined experimentally. Computational modeling techniques can be used to provide the essential structural data needed to shed light on structure-function relationships in TM proteins.

Some of the advanced modeling approaches that can help resolve the unique challenges encountered in predicting the 3D structure of α-helical TM proteins will be presented. The usefulness of standard homology-modeling procedures is limited because the number of available TM-protein structures is small. In many cases, moreover, it is difficult to align the sequences of the query and the template proteins because of the weak sequence similarity between them. Additional ways to predict the location of TM helices in the polypeptide chain, by employing fold recognition, hydrophobicity scales, or other tools, may be helpful in improving the alignment accuracy.

When a structural template is not available, low-resolution electron-density maps, obtained from cryo-electron microscopy (cryo-EM) or preliminary X-ray diffraction studies, can be used to restrict the search in conformational space. At the correct resolution, the locations of TM helices can be roughly determined even when the amino acids are not visible. When these data are combined with physicochemical characteristics of amino acids (such as their hydrophobicity) and with evolutionary conservation analysis of the protein family, the location of the amino acids can be modeled.

After modeling, it is imperative to assess the quality of the structure and estimate the level of confidence of the prediction. To this end, it is often helpful to estimate the evolutionary conservation of the amino acids and project them onto the model-structure (e.g., Fig. 1). The expectation is that the protein core and functional regions (such as narrow channel pores and ligand-binding sites) will accommodate evolutionarily conserved amino acids while the periphery will be more variable. Deviations from this pattern might reflect inaccuracies in the model. Furthermore, because X-ray crystal structures of TM proteins are often determined on the basis of electron-density maps of limited resolution, it might be useful to examine their evolutionary profiles as an independent measure of their validity.

Figure 1. Conservation analysis of erroneous and correct structures of ABC transporters. The retracted structure of MsbA (panel A) and the structure of sav1866 (panel B) are colored according to conservation, using the color scale of the ConSurf webserver (http://consurf.tau.ac.il). Turquoise-through-maroon indicates variable-through-conserved. The most highly conserved residues, receiving grades of 8 or 9, along with the outermost variable (grades of 1 or 2) are shown as spheres. The two upper panels show a side view of the two proteins with their cytoplasmic sides facing down. Predicted membrane boundaries are shown in grey. The nucleotide-binding cytoplasmic domains of both MsbA and sav1866 were omitted for clarity. In both proteins, the cytoplasmic ends are highly conserved and form contacts with the cytoplasmic domains. The two lower panels show a top (and closer) view of the same proteins. Conserved residues for sav1866 are evidently buried in the protein core while variable residues face the lipids, as anticipated. However, the evolutionary profile of the retracted structure of MsbA shows a different pattern: some conserved residues face the lipids and some variable residues are buried in the core.

K07. Paradigm Shifts and Privileged Structures in Drug Discovery

Klaus Müller. F. Hoffmann-La Roche Ltd Pharmaceutical Research - Science & Technology Relations

During only a few decades, drug discovery has undergone dramatic changes spurred by the emergence of new technologies and novel concepts. There is no end to this process.

The surge of molecular biology and biotechnology resulted in a paradigm shift from chemical intuition-dominated to molecular bio-mechanism-based approaches in drug discovery. The firm establishment of structural biology also in industrial environments together with much development in theoretical concepts provided the 3D-structural basis for ever more efficient molecular design. Miniaturization, parallelization, and automation technologies put high-throughput compound handling and screening on qualitatively and quantitatively new levels, and resulted in the introduction of screening technologies
for bio-physicochemical and *in vitro* pharmacological compound properties. The latter produced a dramatic paradigm shift of the drug discovery process and the emergence of multi-dimensional lead optimization much beyond potency and selectivity.

After a decade of hype and unfulfilled hopes regarding combinatorial and high-throughput chemistry, drug discovery now relies on fragment screening, small motif-based compound libraries, molecular scaffold design, and privileged structures.
<table>
<thead>
<tr>
<th>ID</th>
<th>Representative figure</th>
<th>Title</th>
<th>Authors</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td>From protein local structure prediction to local flexibility prediction: flexibility versus structural prediction errors</td>
<td>Aurelie Bornot, Catherine Etchebest and Alexandre G. de Brevern</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Solvent in protein-protein interactions and its impact on protein contacts prediction</td>
<td>Sergey Samsonov, Joan Teyra, Gerd Anders and M. Teresa Pisabarro</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Discovery of novel chemokines with automated structure-based functional protein annotation methods</td>
<td>Aurelie Tomczak and Maria Teresa Pisabarro</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Sasichandran: A sequence specific dihedral angle database and evaluation tool</td>
<td>S. M. Minhas Ud-Dean and Mahdi Muhammad Moosa</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Characterisation of novel proteins involved in mitosis by structure-based computational methods</td>
<td>Jana Sontheimer, Mirko Theis, Frank Buchholz and Maria Teresa Pisabarro</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>How does the quality of homology models affect the accuracy of ligand-protein docking results?</td>
<td>Annalisa Bordogna, Alessandro Pandini and Laura Bonati</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Automated Scaffold Selection in Enzyme Design</td>
<td>Christoph Malisi, Oliver Kohlbacher and Birte Hoecker</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Comparison of proteins flexibilities by self-organizing maps: the test case of SH3 and its mutants</td>
<td>Domenico Fraccalvieri, Alessandro Pandini, Fabio Stella and Laura Bonati</td>
<td>22</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Information about solvent contacts improves foldability in knowledge-based protein models</td>
<td>Sanne Abeln and Daan Frenkel</td>
<td>23</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>FiberDock: Modeling Restricted Backbone Flexibility in Docking</td>
<td>Efrat Mashiach, Ruth Nussinov and Haim J. Wolfson</td>
<td>24</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>A density based approach to define structural alphabets</td>
<td>Alessandro Pandini and Jens Kleinjung</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>Architectures and Functional Coverage of Protein Interfaces</td>
<td>Nurcan Tuncbag, Attila Gursoy, Ozlem Keskin and Ruth Nussinov</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Insights into conformation change on protein docking using normal mode simulations</td>
<td>Sara Dobbins, Victor Lesk and Michael Sternberg</td>
<td>27</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>Sequence and Structural Features of Enzymes and their Active Sites by EC Class</td>
<td>Tracey Bray, Andrew Doig and Jim Warwicker</td>
<td>28</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>Three-Dimensional Structural Determinants of Amino Acid Conservation in Proteins</td>
<td>Sungsam Gong and Tom Blundell</td>
<td>30</td>
</tr>
<tr>
<td>Page</td>
<td>Title</td>
<td>Authors</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Structural rearrangement in the TCRpMHC formation in reaction to agonistic and antagonistic peptides</td>
<td>Bernhard Knapp and Wolfgang Schreiner</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Fractal Analyses of Proteins: Verifying an Equation of State</td>
<td>Marina de Leeuw [1], Shlomi Reuveni [2], Joseph Klafter and Rony Granek</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Ab initio modelling: from Holy Grail to everyday protein bioinformatics tool</td>
<td>Daniel Rigden</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Genome sequencing delivers diminishing returns for homology detection</td>
<td>Daniel Chubb, Lawrence Kelley and Michael Sternberg</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Homology modelling based protein functional residues prediction</td>
<td>Raquel Minardi and François Artiguenave</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Effective atomic interactions for the characterization of protein-ligand binding interfaces</td>
<td>Alex Slater, Evandro Ferrada and Francisco Melo</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Human-curated database of three-dimensional protein-DNA complexes</td>
<td>Tomas Norambuena and Francisco Melo</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Prediction of Functional Residues in Protein Structures with Insights into Protein Phosphorylation</td>
<td>Vladimir Vacic [1], Lilia Iakoucheva [2], Stefano Lonardi [3] and Predrag Radivojac</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Pore-Walker: a Novel Tool for the Identification and Characterization of Channels in Transmembrane Proteins from Their Three-Dimensional Structure</td>
<td>Marialuisa Pellegrini-Calace, Tim Maiwald and Janet M Thornton</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>The effect of loops on the structural organization of ( \alpha )-helical membrane proteins</td>
<td>Hagai Meirovitch, Oznur Tastan and Judith Klein-Seetharaman</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>iMembrane: Homology-Based Membrane Insertion of Proteins</td>
<td>Sebastian Kelm, Jiye Shi and Charlotte M. Deane</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods</td>
<td>Stefano Lise, Cedric Archambeau, Massimiliano Pontil and David Jones</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Assisted crystallographic RNA model building: A directed rotameric approach for building the RNA backbone</td>
<td>Kevin Keating and Anna Marie Pyle</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Predicting stable regions in membrane proteins: a novel coarse-grained approach applied to bacteriorhodopsin</td>
<td>Anne Tuukkanen, Frank Dressel, Annalisa Marsico, Rainer Winnenburg, Dirk Labudde and Michael Schroeder</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>A novel approach to structural modelling of large protein complexes</td>
<td>Anne Tuukkanen, Andreas Henschel and Michael Schroeder</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Determining protein-ligand ‘interaction space’ in the druggable portion of Mycobacterium tuberculosis proteome</td>
<td>Kalidas Yeturu and Nagasuma Chandra</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Virosphere specific protein folds</td>
<td>Aare Abroi</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Assessing conformational diversity in proteins using evolutionary information</td>
<td>Nicolas Palopoli, Ezequiel Juritz, Sebastian Fernandez Alberti and Gustavo Parisi</td>
<td>45</td>
<td></td>
</tr>
</tbody>
</table>
Computational Characterization of Human Serum Paraoxonase’s Structure and Its Interaction with VX  

Wenling Chang, Steven Fairchild, Matthew Peterson, Chang-Guo Zhan, Adel Hamza and Douglas Cerasoli

Structure-based prediction of antibody epitopes  

Petr Ponomarenko, Ruben Abagyan, Philip Bourne and Julia Ponomarenko

FREAD revisited: template-based loop prediction  

Yoonjoo Choi and Charlotte Deane

Assessing the Stability of Protein Complexes within Large Assemblies  

Frederic Cazals and Tom Dreyfus

A comparative structural analysis of aquaporins: Pore-lining residues and electrostatics of the channel  

Romina Oliva [1], Giuseppe Calamita [2], Janet M. Thornton [3] and Marialuisa Pellegrini-Calace

Molecular evolution of class A protein-coupled receptors probed by the proline pattern of transmembrane helix 2  

Julie Devillé, Julien Rey and Marie Chabbert

Prediction of causative effects of disease-related mutations at the molecular level by integrated bioinformatic analyses  

Jan Kosinski and Janusz Bujnicki

STP: A Program to Predict Protein Binding Surfaces  

Wissam Mehio [1], Graham J.L. Kemp [2], Paul Taylor [1] and Malcolm D. Walkinshaw

Organic enzyme cofactors – from structure to function  

Julia Fischer, Gemma Holliday and Janet Thornton

Protein motion and visualization using game engine  

Maria Francesca Zini, Tiziana Loni, Raluca Andrei, Stefano Cianchetta, Marco Callieri and Monica Zoppe'

Bioinformatics for GPCRs oligomerization  

Wataru Nemoto, Kazuhiko Fukui and Hiroyuki Toh

Genome-wide structural modeling of protein-protein interactions  

Petras Kundrotas, Zhengwei Zhu and Ilya Vakser

Conformational changes in redox pairs of protein structures  

Merridee A Wouters, Richard A George, Naomi L Haworth, Lina L Feng, Jason Y Liu and Samuel W Fan

Using a Differential Geometry Approach for Characterizing Biological Relevant Interfaces  

Shula Shazman Gershon Elber and Yael Mandel-Gutfreund

Protein sequence and structure optimisation in one probabilistic framework  

Gundolf Schenk and Andrew Torda

A new protein binding pocket similarity measure based on comparison of 3D atom clouds: application to ligand prediction  

Brice hoffmann, Mikhail Zaslavskiy, Jean-Philippe Vert and Véronique Stoven

Multiple Alignments and Phylogenies of Protein Structures  

Thomas Margraf and Andrew Torda
<table>
<thead>
<tr>
<th>Page</th>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>3D-BLAST: Towards fast 3D protein shape superposition, comparison, and classification</td>
<td>Lazaros Mavridis and David W. Ritchie</td>
<td>61</td>
</tr>
<tr>
<td>55</td>
<td>Alpha-helical transmembrane protein fold prediction using residue contacts</td>
<td>Timothy Nugent and David Jones</td>
<td>62</td>
</tr>
<tr>
<td>56</td>
<td>The overlap of small molecule and protein binding sites within families of protein structures</td>
<td>Fred P. Davis and Andrej Sali</td>
<td>63</td>
</tr>
<tr>
<td>57</td>
<td>Structural Signature of Antibiotic Binding Sites on the Ribosome</td>
<td>Hilda David-Eden and Yael Mandel-Gutfreund</td>
<td>63</td>
</tr>
<tr>
<td>58</td>
<td>A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation</td>
<td>Sebastien Loriot, Frederic Cazals, Michael Levitt and Julie Bernauer</td>
<td>64</td>
</tr>
<tr>
<td>59</td>
<td>A 3D image database from structural to cellular and systems biology: The next-generation Electron Microscopy Data Bank</td>
<td>Christoph Best, Cathy Laswon, Richard Newman, Glen van Ginkel, Helen Berman and Kim Henrick</td>
<td>64</td>
</tr>
<tr>
<td>60</td>
<td>Automated Ligand-based Active Site Alignment</td>
<td>Abraham Heifets and Ryan Lilien</td>
<td>65</td>
</tr>
<tr>
<td>61</td>
<td>A graphical model approach for predicting free energies of association for protein-protein interactions under backbone and side-chain flexibility</td>
<td>Hetunandan Kamisetty, Chris Bailey-Kellogg and Christopher Langmead</td>
<td>66</td>
</tr>
<tr>
<td>62</td>
<td>FlexPepDock: Accurate modeling of peptide-protein complexes with full peptide flexibility</td>
<td>Ora Furman (Schueler), Nir London and Barak Raveh</td>
<td>68</td>
</tr>
<tr>
<td>63</td>
<td>How good can template-based modelling be?</td>
<td>Braddock K. Lance, Graham Wood and Charlotte Deane.</td>
<td>69</td>
</tr>
<tr>
<td>64</td>
<td>Ligand-based and structure-based virtual screening to identify potential inhibitors of drug-efflux transporters</td>
<td>Sergio Mares-Samano, Raj Badhan and Jeffrey Penny</td>
<td>70</td>
</tr>
<tr>
<td>65</td>
<td>Simulation of induced structural transitions in an isolated Kv1.2/2.1 voltage-sensor important for the gating mechanism</td>
<td>Christine S. Schwaiger, Pär Bjelkmar and Erik Lindahl</td>
<td>71</td>
</tr>
<tr>
<td>66</td>
<td>Expert knowledge enhanced structure based profile HMMs for protein sequence families</td>
<td>Stefan Wegenkittl, Felix Auer, Daniel Bindreither and Peter Lackner</td>
<td>72</td>
</tr>
<tr>
<td>67</td>
<td>Quality Assessment Testset for Structure Alignments</td>
<td>Christoph Berbalk and Peter Lackner</td>
<td>72</td>
</tr>
<tr>
<td>68</td>
<td>Shedding light on transcription factor binding sites using molecular dynamics</td>
<td>Nina M. Fischer, Dierk Wanke and Oliver Kohlbacher</td>
<td>73</td>
</tr>
<tr>
<td>69</td>
<td>Automated classification of enzymes using the Active Site Classification (ASC) method</td>
<td>Marc Rötting and Oliver Kohlbacher</td>
<td>74</td>
</tr>
</tbody>
</table>
TopDomain-web: interactive protein domain decomposition

Stefan Senn and Manfred J. Sippl

COPS - A Novel Workbench for Explorations in Fold Space

Stefan J. Suhrer and Manfred J. Sippl

COPS-Benchmark: a structure-derived benchmark for homology and fold detection methods

Markus Gruber, Karl Frank and Manfred Sippl

An analytical solution for finding voids and bottlenecks within macromolecules

Fredrik Boulund, Viktor Jonsson, Erik Sterner and Graham Kemp

Interface flexibility in hub proteins

Arianna Forni and Franca Fraternali

QMEANfamily: Integration of evolutionary information in the composite scoring function QMEAN

Pascal Benkert and Torsten Schwede

ModLink+: improving fold recognition by using protein-protein interactions

Oriol Fornes, Ramon Aragues, Jordi Espadaler, Jaume Bonet, Marc A. Marti-Renom, Andrej Sali and Baldo Oliva

Structural and Functional Roles of Fragments of the Ankyrin Repeat Protein p18INK4c

Petr Sklenovsky and Michal Otyepka

The explicit water models in AMBER force field affect the stability of the artificial peptide chignolin in molecular dynamic simulations

Petra Florova, Petr Sklenovsky and Michal Otyepka

The Protein Model Portal

Lorenza Bordoli, Konstantin Arnold, Florian Kiefer, Jürgen Haas, Michael Künzli, James Battey, Michael Podvines, John Westbrook, Helen Berman and Torsten Schwede

Extension, distribution and causes of conformational diversity in proteins

Ezequiel Juritz, Sebastian Fernandez Alberti and Gustavo Parisi

Similarity of Chemical Mechanisms in Functionally Analogous Enzymes

Daniel E. Almonacid, Emmanuel R. Yera, John B. O. Mitchell and Patricia C. Babbitt

Quantum States of Biomolecular Breathers. Insights from Semiclassical and Feynman Path Integral Quantization of Protein Solitons

Alexander Kantardjieff
ABSTRACTS

3. From protein local structure prediction to local flexibility prediction: flexibility versus structural prediction errors.

Aurelie Bornot, Catherine Etchebest and Alexandre G. de Brevern. INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques, Université Paris Diderot - Paris 7, France.

In this study, we addressed the question of the structural “predictability” of a protein sequence with regards to its structural flexibility properties. Protein local structure dynamics were analyzed using X-ray experiments and molecular dynamics simulations. Finally, an original flexibility prediction method from protein sequence was proposed.

In this study, we addressed the question of the structural “predictability” of a protein sequence with regards to its structural flexibility properties. Protein local structure dynamics were analyzed using X-ray experiments and molecular dynamics simulations. Finally, an original flexibility prediction method from protein sequence was proposed.

Protein flexibility is critical for biological processes. Therefore, toward the understanding of protein function at the molecular level, structural analysis need to take into account flexibility information. We previously elaborated a library of 120 overlapping representative fragments of 11 amino-acids, encompassing all known protein local structures. Moreover, we developed a local structure prediction method and defined different confidence indexes for directly assessing the prediction quality [1,2]. A very satisfying prediction rate was obtained. Nevertheless, some sequence regions remained more difficult to predict.

In this study, we went further and addressed the question of the structural “predictability” of a sequence with regards to its structural flexibility properties inside protein structures. We analysed local structure flexibility features in proteins by relying on: (i) B-factors from X-ray experiments and (ii) backbone fluctuations in solution observed in molecular dynamics simulations. Using the complementarities of both data sources, specific local structure dynamic properties but also correlations between flexibility degree and local structure prediction rates were enlightened.

Finally, based on these structure-flexibility relationships and on our local structure prediction method, an original flexibility prediction method from sequence was developed. Three classes, rigid, intermediary and flexible, were predicted with a well balanced prediction rate of 49% and only 6% of confusion between the two extreme classes. When adapted for comparison with the cutting-edge two-class prediction method PROFbval [3], our strategy was proved very competitive with 75 and 85% of accuracy and coverage respectively.


Sergey Samsonov, Joan Teyra, Gerd Anders and M. Teresa Pisabarro. BIOTEC TU Dresden, Germany.

Although solvent is important in protein interfaces, its impact on protein-protein interactions in computational studies is often ignored. We analyze water-mediated interactions in protein interfaces by a MD approach and, in addition, demonstrate that introduction of solvent into the correlated mutations concept improves to a certain extent protein contacts predictions.

Water constitutes the cellular environment for biomolecules to interact. Solvent is important for protein folding and stability, and it is also known to actively participate in many catalytic processes in the cell. However, solvent is often ignored in molecular recognition and not taken into account in protein-protein interaction studies and rational design.

The goals of our studies have been to 1. use a MD approach to characterize dynamic and energetic properties of wet spots (interfacial residues interacting only through one water molecule) and the water molecules forming them; 2. analyze the impact of including solvent for protein contacts prediction.

Our MD data shows that water molecules can play an important
role in interaction conservation in protein interfaces by allowing sequence variability in the corresponding binding partner. We find that disregarding interfacial solvent may cause inaccuracies in the application of correlated mutations based approaches in the complete analysis of protein interfaces and the prediction of protein interactions. The correlated mutations concept is based on the assumption that interacting protein residues coevolve, so that a mutation in one of the interacting counterparts is compensated by a mutation in the other. We use linear combinations of the predictions obtained by the application of two different similarity matrices: a standard “dry” similarity matrix (DRY) and a “wet” similarity matrix (WET) derived from all water-mediated protein interactions in the PDB. We analyze two datasets containing 50 domains and 10 domain pairs from PFAM and compare predictions obtained by using a combination of both matrices. Our results, despite the complexity of their possible general applicability, open up that the consideration of water may have an impact on the improvement of the contact predictions obtained by correlated mutations approaches.


Aurelie Tomczak and Maria Teresa Pisabarro. BIOTEC TU Dresden, Germany.

We developed an automated framework for structure-based annotation of proteins and use it to identify novel chemokines, which are proteins essential for guiding immune cell migration. We integrate publicly available sequence and structure information and feature prediction tools with fold recognition and identified promising candidates that are currently experimentally validated.

Classic bioinformatics approaches based on sequence similarity are often useful to find homologous proteins and thus infer the function of unknown ones. In case sequence-based annotation methods fail, the application of structure-based methods (i.e. Fold recognition) can help to confirm tentative functional assignments. These methods have already proven to be successful for individual proteins but, due to the large amount of data obtained by high-throughput approaches, automation of structure-based functional annotation methods is needed.

We have developed a framework for automatic structure-based annotation of proteins and use it to identify novel chemokines, which are small secreted signal proteins guiding immune cell migration and playing an important role in processes like angiogenesis, haematopoiesis, HIV and tumor growth. We integrated data and tools from different sources: public databases containing experimental structural and sequence data (i.e. PDB, Swiss-Prot), sequence feature prediction tools (signal peptide, transmembrane, secondary structure prediction), structure-based computational methods and in-house developed methods describing structural and functional features of the chemokine fold family (3D-descriptors of disulfide bonds and fold coverage). We extracted a large dataset of uncharacterized protein sequences containing cysteine residues from public databases and screened it with our automated framework to discover novel chemokines with remote sequence identity.

We found several putative novel chemokines and detailed 3D models have been built for them. Their stability and energetic properties have been characterized using computational techniques such as molecular dynamics simulations, and promising candidates are currently in the process of being confirmed experimentally.


S. M. Minhaz Ud-Dean and Mahdi Muhammad Moosa. University of Dhaka, Bangladesh.

Estimation and prediction of dihedral angle angle can be used to validate both theoretically predicted and experimentally determined structures. This idea was used to develop sequence specific dihedral angle prediction tool, Sasichandran. This tool can also evaluate a protein structure using information of sequence specific distribution of Ramachandran angles.

Availability: http://sasichandran.8bhuj.com

Protein structure prediction remains one of the greatest unsolved challenges of computational biology. Structure evaluation is used both for assessing the quality of the newly determined structures and predict the structure of newly sequenced proteins. Massive amount of sequence data generated by several ongoing genome sequencing projects is making protein structure and function prediction even more important.

Ramachandran et al. introduced the concept of allowed dihedral angles (Phi, Psi) based on stereochemical hindrance as a parameter to describe protein backbone structure or secondary structure (Ramachandran, et al., 1963). The plot of these angles has been extensively used in determining protein structure (Kleywegt and Jones, 1996; Morris, et al., 1992), in defining protein secondary structure folds (Chou and Fasman, 1974; Ho, et al., 2003; Munoz and Serrano, 1994) and in the verification of experimentally determined protein structures (Hovmöller, et al., 2002; Laskowski, et al., 1993; Morris, et al., 1992).

Although Dayalan et al. (Dayalan, et al., 2006) developed a database of Phi-Psi dihedral angles of different amino acid combinations
(5,227 non-redundant high resolution protein structures) they did not provide any option for prediction or evaluation of backbone dihedral angle. Conformational Angles DataBase (CADB) of proteins by Sheik et al. (Sheik, et al., 2003) also provides a similar database of Phi-Psi angles of 6146 protein chains. But neither provides option for prediction or evaluation of protein structures. Our tool provides options for sequence specific dihedral angle data retrieval and evaluation of structures based on provided data.

Database: PDB files of proteins with resolution better than 1.5 Angstrom were downloaded and sequences with greater than 95% homology were discarded. Non-standard amino acids were excluded by splitting the polypeptide chains into two fragments whenever they were encountered. Phi and Psi values for each amino acid were calculated. For all possible two, three and four amino acid combinations respective values of Phi and Psi were computed. For each of the Phi- Psi combinations of each AA combination, Phi-Psi instances within 10 degree (both Phi and Psi) were counted. The maximal Phi-Psi combinations within the acceptable range were considered most probable. The probability (in percent) of most probable combination was computed by dividing the number of instances within the range with the total number of instances.

Implementation: The online version of the database includes an interface for calculation of dihedral angles from pdb files, an interface for extracting the most probable dihedral angles of a given sequence and appraisal of a protein structure in pdb format. Similar options are also available in command line based offline version. The offline version can be run in batch mode for multiple files.

The dihedral calculation from pdb files is carried out by a script adapted from Carroll (Carroll, 2005). Most probable dihedral angle is retrieved by searching the subsequences of the query in the database of most probable dihedral angles. The appraisal tool first calculates the dihedral angles of a given pdb file and then counts the number of Phi- Psi instances within 10 degree (both phi and psi) of the Phi- Psi combination for each of the given residues and respective probabilities are determined based on the frequency of appearance.

Sasichandran Dihedral Angle Extractor: Prediction of most probable two, three or four amino acid blocks can be carried out by going to respective pages; an option of exhaustive search is also available. Figure shows an example output of exhaustive search. The offline version works on the same principle. In the command line one can enter the intended peptide sequence and then select the mode (exhaustive, 2AA, 3AA or 4AA). The result is both displayed on the screen and saved in a .tsv file.

Ramasekharan: This tool calculates dihedral angle of structures from PDB database (Berman, et al., 2000). PDB ID of respective files should be submitted to calculate Phi-Psi dihedral angles. The offline version would take a folder containing .pdb files and return .tsv files containing Phi-Psi angles for each of the standard amino acid residues. Additionally the calculated angles are also displayed on the command line.

Sasichandran Dihedral Angle Appraiser: The tool gives the option of appraising dihedral angle. This can be done by uploading protein structure file in pdb format or pasting the formatted file in the input window. The tool first calculates dihedral angles of each amino acid residues. The input is then split into overlapping ten amino acid fragments. The probability of each observed dihedral angle is calculated for each fragment and given as output.

REFERENCES

7. Characterisation of novel proteins involved in mitosis by structure-based computational methods.


With full sequenced genomes, computational methods are gaining importance in characterisation of novel proteins. We have developed a framework for automatic structure-based function annotation and used it for analysing phenotypic data from genome-scale RNAi screenings, which has resulted in characterisation of a novel mitotic protein regarding its structure and function.

The function of many proteins is still unknown though the amount of protein-related data from genomics, proteomics and other high-throughput-screening methods is increasing rapidly. Classical sequence-based bioinformatics approaches often fail to predict protein function, in case no homologous protein is found. As protein
structure is more conserved than sequence in protein families, structure-based methods (e.g. fold recognition) may recognize possible structural similarities even at low sequence similarity and therefore provide information for function inference. These methods have already been proven to be successful for individual proteins but, due to the large amount of data, an automated procedure of structure-based functional annotation of proteins is needed.

With the aim of high-throughput analysis we have developed a computational framework for functional automated structure-based annotation of uncharacterised proteins: We apply fold recognition methods to derive potential protein functions and validate our findings by combining structural and functional information from publicly available resources such as PDB, SwissProt or SCOP.

We have used our platform to analyse data from human genome-scale RNAi-screenings in order to link phenotype to molecular function. We have been focussed on hits corresponding to uncharacterised proteins from a cell viability RNAi screen. For fold recognition we have used a foldlibrary containing all known 3D-structures annotated with GO-terms related to the RNAi phenotype. This has resulted in the identification of a structural motif known to interact with a component of the mitotic spindle-assembly checkpoint.

8. How does the quality of homology models affect the accuracy of ligand-protein docking results?


We investigated how the quality of homology models affects the accuracy of ligand-docking experiments. Our results demonstrated that the quality of the modelled binding site, assessed by comparison to the native or the template protein structure, is the most informative measure to predict the accuracy of the docking results.

Motivations:
For an increasing number of cases, molecular docking calculations require the employment of 3D-models derived from homology modelling techniques. Therefore predicting the accuracy of the docking results on the basis of an a priori evaluation of the model quality is of great interest. Aim of this work is to identify quantitative relationships between indexes of model quality and the accuracy of binding geometries obtained by molecular docking calculations.

Methods:
A representative subset of the CCDC/Astex Test Set (1) was chosen as the reference set of X-ray structures. This is a large and diverse collection of known protein-ligand complexes extensively used to validate docking methods.

To generate a corresponding dataset of models, identification of candidate templates was performed by sequence similarity search using PSI-BLAST (2). Templates covering a broad range of sequence identity (15-85%) were selected to provide a reliable sampling of different evolutionary distances. Additionally, to provide a wider spectrum of model quality, for each case, modelling was performed by employing two methods, the automatic server I-TASSER (3) and Modeller 9v1 (4), and the reference alignments for Modeller were generated by using different approaches. T-Coffee (5) was used for mono- and multi-template sequence alignments, Praline (6) for profile-profile alignments, and TMalign (7) for structural alignments. The final dataset included 245 models and 21 X-ray structures.

In a first step, the quality of the models was assessed by means of indexes derived from direct comparison to the known native structures: three structural alignment approaches implementing both global- and local-fit algorithms were employed (Dali, LGA, ProFit) (8, 9, 10). These “calculated” indexes provided a direct measure of conformity to the target.

Furthermore, a group of predictive indexes of model quality derived without any reference to the known native structures: three structural alignment approaches implementing both global- and local-fit algorithms were employed (Dali, LGA, ProFit) (8, 9, 10). These “calculated” indexes provided a direct measure of conformity to the target.
AutoDock (14) was used to perform docking calculations. The accuracy of results was assessed by distance Root Mean Square Deviation (dRMSD): ligand-active site distances in the model were compared with the corresponding ones in the X-ray structure. This provided a reliable measure that is less affected by small changes in the active site geometry.

Results:

To account for possible differences in docking performances on specific classes of ligands, the analysis was performed on the whole dataset as well as on “homogeneous” subsets composed by proteins that share the same fold and bind the same class of ligands.

A first aim of this work was to investigate the correlations between model quality indexes calculated by direct comparison with the native structure and the docking accuracy. For the whole test set, that includes a broad and diverse collection of protein structures, we found that the accuracy of binding geometries obtained by molecular docking are mostly dependent on the ability of the model to correctly reproduce the active site geometry. Indeed, the binding site RMSD and the Atomic Contact Score (15) showed the highest correlation with the dRMSD. A multivariate analysis on all the calculated indexes confirmed the site RMSD as the most relevant one, highlighting a minor importance of indexes of global quality in the regression model.

Interestingly, for subsets including proteins with similar structures and similar ligand binding sites, strong and statistically significant correlations between the accuracy of docking results and all these quality indexes were found.

The most ambitious aim of this work was to predict the accuracy of docking results from those indexes of model quality commonly used to validate protein structures. Indeed, we showed that some of the “predicted” indexes do exhibit a clear relationship with the accuracy of docking geometries. The analysis on the complete test set highlighted the importance of geometrical coherence of the modelled binding site with that of the template structure in providing accurate docking results.

In this case, the best multivariate regression model emphasized the role of both local and global quality indexes. This result was also confirmed by the analysis of the “homogeneous” subsets.

In conclusion, for the first time we provided quantitative relationships between “calculated” indexes and docking results and we demonstrated that further analyses on “predicted” indexes could lead to the identification of general criteria for an a priori prediction of the accuracy in ligand-protein experiments. Moreover, we presented evidence that accuracy prediction is already feasible for proteins of the same fold: by performing docking calculations on models whose experimental structure in the bound state is known, the accuracy of docking results can be easily predicted also for homologous proteins with unknown structure.

References:


We present an algorithm for searching attachment sites for catalytic motifs in protein structures. It identifies pairs of geometrically suitable backbone positions, and combines these with a graph clique search to complete attachment sites. Performance was tested in a benchmark involving identification and geometric recapitulation of catalytic motifs in a large structure database.
ScaffoldSelection then examines protein structures for potential attachment sites. First, generally eligible backbone positions for attachment are determined. They should be a part of or near a surface pocket-like geometry, and not be in the center of a secondary structural element, because mutations introduced here have a relatively high probability to disturb protein stability. Among these backbone positions, the algorithm looks for pairs of C-alpha-C-beta vectors whose distance and angle values fall within the range of one of the catalytic residue pairs, and are therefore a potential site for the introduction of that pair in the necessary geometry. Now, a compatibility graph is constructed with a node for each encountered C-alpha-C-beta vector pair; two nodes are connected if they assign the same scaffold backbone position to the same catalytic residue. The cliques in this graph that contain exactly one node for each catalytic residue pair represent potential attachment sites for the complete motif. Rotamers of the catalytic residues are then attached to these complete attachment sites and the rotamer set that creates the geometry nearest to the input is determined. These matches are investigated further, e.g., for steric clashes with the scaffold backbone and for space to accommodate the substrate or transition state of the enzymatic reaction. The final output of the algorithm is a ranked list of attachment sites in candidate scaffolds that can then be investigated further.

In order to judge the ability of our method to identify suitable scaffolds and attachment sites for enzymatic or binding-site motifs, we assessed the program’s performance using twelve different motifs that were extracted from naturally existing enzymes. We tested the ability of ScaffoldSelection to recapitulate the geometry of the native enzymatic sites that are taken from x-ray structures solved with a transition state-like inhibitor bound to the active site. As a benchmark for the algorithm, we decided to use the motif test set compiled by Zanghellini et al. (3). In addition to the ten motifs from that set, we included a catalytic triad from subtilisin and a motif extracted from chorismate mutase. Subtilisin was chosen because the catalytic triad motif evolved independently in the enzyme families of subtilases and trypsin-like serine proteases, thereby providing a perfect test scenario to check if with a motif extracted from one family a structure from the other family could be identified as a suitable scaffold. The chorismate mutase reaction was added because it has been the subject of numerous protein engineering studies, and it complements the set by the representative of the mutase EC-class. In order to obtain reliable results on the performance of ScaffoldSelection, we looked for suitable matches of the test motif also in a large database of about 3,500 structures. This database is a non-redundant subset of the whole Protein Data Bank obtained from the NCBI Molecular Modeling Database and was processed further to remove structures that are not suitable as scaffolds, e.g., structures solved to low resolution only. ScaffoldSelection was able to identify all twelve original attachment sites; five of them were ranked in the first percentile of the resulting ranked list. The accompanying figure shows four results: The wild type geometries of the enzymatic motifs including a substrate or inhibitor structure are shown in red, ScaffoldSelection proposed geometries are shown in green. In result C ScaffoldSelection chose one rotamer differing from the wild type, which leads to a less than optimal recapitulation of the substrate position; results A, B and D show excellent recapitulation of the original geometry. The algorithm is able to identify matches in a structure quite fast for motifs containing three residues, using less than a minute per structure in most cases. The method also succeeds in identifying native sites in different x-ray structures of the test set proteins and in homologs of these; and ScaffoldSelection holds up well in comparison to methods attempting a similar goal; it is at least as effective, and significantly faster. Thus, our approach to rapidly identify attachment sites for a specified motif in a large set of protein structures is well suited to preselect scaffolds for computational enzyme design.

REFERENCES


A use of self-organizing maps in the analysis of molecular dynamics data is presented. The method was tested on a group of mutants of the SH3 domain and allowed to easily recover information about differences in the domain flexibilities that affect the biological function.
Introduction
A new and challenging problem in Structural Bioinformatics is the development of accurate methods for comparing protein flexibilities derived from Molecular Simulations. The increasing number of data on protein dynamics can be useful to run comparisons of mutants or proteins belonging to the same family, to observe the influence of limited modifications on the biological function, as well as to analyze functional relationships among evolutionary divergent proteins. While in the former case the comparison could be driven by structural superimposition, in the latter, a set of new techniques should be developed to run comparisons independently of the structural alignment, in order to avoid errors from incorrect alignments. To this purpose in this study we present a use of self-organizing maps (SOM)[1] in the analysis of molecular dynamics (MD) trajectories of different domains to detect similarities and differences in their flexibilities. The method was tested on the Src-Homology region-3 (SH3) domain and a group of its single mutants. The SH3 domain is a small protein module (about 60 residues) that mediates protein-protein interactions in a large variety of cellular processes. Several studies demonstrated the importance of conformational dynamics in defining the binding specificity of SH3 domains. The mutants employed in this study are: R21A, R21G, N47G, N47A, A56G, and A56S. These mutations involved functional and structural change.[2,3]

Methods
For each structure, after a minimization with steepest descent, MD simulations were run using the GROMACS[4] (version 3.3.3) package. All the structures were inserted in a SPC water octahedral box, with periodic boundary conditions. Simulations of 40ns were performed with the Gromos 43a2 force field, in the NPT ensemble, with a time step of 2fs. The extraction of the data on local flexibility was performed after essential dynamics (ED) analysis.[5] This technique, based on principal component analysis (PCA) of conformational ensembles, is aimed to extract informative direction of motion in a multidimensional space.

SOM analysis is a powerful method for visualization and treatment of high dimensional data, based on unsupervised Neural Network. [1] A SOM consists of a bi-dimensional grid of so-called neurons, and a model vector associated. Each model vector represents a group of similar data vectors. A wide set of conformations obtained via MD simulation was used as learning dataset, each conformation being synthesized through a data vector of geometric descriptors. The SOM analysis was performed by using the SOM Toolbox 2.0 for MatLab.[6] The problem of SOM model selection, architecture, preprocessing parameters, learning algorithm and parameters, has been tackled by designing and implementing an experimental design plan. Cluster analysis was performed on the output maps and the characteristics of each cluster were evaluated by analyzing both the representative structures and the intra-cluster RMSF values.

Results
Initially, a comparison of flexibilities was driven by standard sequence and structural alignment followed by annotation of the Root Mean Square Fluctuation (RMSF) for each residue in the equivalent positions. Comparison of the RMSF profiles demonstrated that it is possible to directly correlate the dynamical behavior with experimental information on both stability and functionality of these domains. Furthermore, SOM was used to analyze data vectors obtained by MD trajectories. The analysis of each mutant trajectory confirmed that SOM is able to analyze a large sample of conformations taken from a MD trajectory and to extract clusters representing similar regions in the conformational space. The comparison of the whole SH3 learning dataset confirmed that with this method it is possible to handle data from different simulations. Indeed, the analysis of the domains belonging to each cluster allowed to easily recover information about similarities and differences in the domain flexibilities that affect their biological function.

Future directions involve the extension of SOM analysis to comparison of protein flexibilities among structurally divergent proteins by means of an alternative representations independent of structural alignment.

References:

11. Information about solvent contacts improves foldability in knowledge-based protein models.


Knowledge-based statistical pair potentials capacitate protein-folding simulations of designed model proteins. However simultaneous simulation of multiple proteins with such potentials often results in large aggregates. We show that, by using a pair potential that includes solvent interactions, the designed proteins remain soluble at temperatures below the folding transition.

Motivation
Statistical pair potentials have successfully been used to simulate
protein folding with designed proteins using simplified structural models. The interaction energies between amino acids for these potentials may be derived from statistical analysis of experimental protein structures in the Protein Data Bank (PDB). The resulting pair potentials have also been used in structure prediction, e.g. for threading methods.

In their natural cellular environment folded proteins must not aggregate. Aggregates may be cytotoxic or compromise the biological function of the proteins. Evolutionary pressure generally ensures that proteins do not aggregate in their natural biochemical environment.

However, using simple pair potentials between amino acids to design foldable model proteins, leads to model proteins that aggregate at the folding temperature when multiple proteins are simulated simultaneously. We show that by including solvent interactions in the pair potentials the designed proteins remain soluble at their folding temperatures and below.

Methods
Pair potentials between amino acids can be established by considering the number of pairs of amino acids that are proximate in experimental protein structures. The interaction energies between two amino acids may be estimated by considering the ratio between the number of observed proximate amino acid pairs and the number of expected pairs based on amino acid abundance. Here, this method is extended to include explicit solvent contacts, creating additional interaction energies for each amino acid type with water. The number of water contacts per residue is based on the surface accessible surface area in the PDB structures. The contact energies are calculated based on the number of observed and expected contacts with water.

To test the potential for the simulation of multiple refoldable proteins, we use a simple lattice model where the directionality of the side chains is taken into account. The solvent interactions can be included through interactions with empty lattice sites. Foldable proteins are designed using a Monte Carlo algorithm that optimises the interaction energies by changing the sequence, while keeping the structure fixed in a compact conformation (Coluzza & Frenkel 2003).

Results
Large aggregates are observed at the folding temperature, when multiple proteins are simulated simultaneously with a pair potential based on amino acid interactions alone (e.g. Betancourt & Thirumalai 1999). This holds both for Monte Carlo simulations with the conventional cubic lattice model, as for simulations with a cubic lattice model in which side chain directionality is included.

When foldable proteins are designed and simulated with the pair potential that includes the solvent, the proteins do not aggregate at temperatures significantly lower than the folding temperature. These simulations were performed with a cubic lattice model that includes side chain directionality. The proteins, designed with the potential that contains solvent interactions, also show better defined hydrophobic cores.

Significance
Protein aggregation, in particular amyloid formation, is associated with several neurodegenerative disease. There is therefore a high interest in modelling the early aggregation stages of amyloid forming proteins. However, under physiological conditions most proteins should not aggregate; when modelling aggregation or amyloid formation, first an appropriate model of non-aggregating proteins should be established. The pair potential presented here allows for qualitatively realistic simulations of the aggregation behaviour in protein systems.

The calculation of the potential is simple to implement, and the concept of adding solvent interactions is easily adaptable to more complex interaction potentials and more detailed protein structure models. Moreover, the proteins designed with the potential that includes solubility have a better defined hydrophobic core. This suggests that it may be useful to include solvent-amino acid interactions in pair potentials for structure prediction, for example for threading methods.


FiberDock is a novel docking refinement method that models backbone and side-chain flexibility. The method minimizes backbone conformations along the most relevant normal modes, which correlate with chemical forces. It uses both low and high frequency modes and therefore is able to model global and local conformational changes.

FiberDock Refinement Result

FiberDock is a novel docking refinement method that models backbone and side-chain flexibility. The method minimizes backbone conformations along the most relevant normal modes, which correlate with chemical forces. It uses both low and high frequency modes and therefore is able to model global and local conformational changes.

Introduction
Upon binding, proteins undergo conformational changes that include both side-chains and backbone movements. These changes often prevent rigid-body docking methods from predicting the 3D structure of a complex from the unbound conformations of its proteins. Handling protein backbone flexibility is a major challenge for docking methodologies, since backbone flexibility adds a huge number of degrees of freedom to the search space. Normal mode
analysis is a commonly used method for analyzing the flexibility of a protein, given a single 3D structure, such as the unbound conformation. The analysis provides a set of possible movements, called normal modes, of the protein backbone, and their vibration frequencies. The lowest frequency normal modes usually describe the large conformational changes a protein can undergo, while high frequency modes describe subtle changes. Many docking methods model backbone flexibility by using only few modes, which have the lowest frequencies (Lindahl & Delarue, 2005; May & Zacharias, 2008). However, studies show that due to molecular interactions, many proteins also undergo small-scale conformational changes, which are described by high frequency normal modes (Petrone & Pande, 2006). Thus, high frequency normal modes should be taken into account in flexible docking refinement methods.

Here we present FiberDock, a new method for flexible docking refinement. The method allows both backbone and side-chain flexibility. It minimizes the structural conformations of the interacting proteins and optimizes their rigid-body orientation. The side-chain flexibility is modeled by a rotamer library and the backbone flexibility is modeled by an a-priori unlimited number of normal modes. The method iteratively applies the most relevant normal mode on the flexible protein. The relevance of a mode is calculated according to its correlation with the chemical forces applied on each atom.

Unlike previous backbone refinement methods, our method uses both low and high frequency normal modes. Hence it is able to model both global and local conformational changes, such as opening of binding sites and loop movements. The method was tested on a benchmark and the results show that the incorporation of backbone flexibility in the refinement process considerably improves the accuracy and the ranking of models of protein-protein complexes.

The FiberDock Method

The FiberDock method refines rigid-docking solution candidates and re-ranks them in order to identify near native models. The refinement takes into account both backbone and side-chain flexibility. The method combines a novel NMA based backbone refinement with our previously developed side-chain optimization and rigid-body minimization method, FireDock (Andrusier el al., 2007). Currently, the backbone flexibility is modeled only for one of the proteins, called the receptor, and the backbone of the other protein, the ligand, is kept rigid.

Normal mode analysis is performed in a pre-processing stage. In this stage the normal modes of the receptor are calculated using the Anisotropic Network Model. The FiberDock algorithm, which is applied on each rigid-body solution candidate, includes four main stages. In the first stage the interface side-chain conformations of both proteins are optimized. The flexibility is modeled by a rotamer library and the optimal combination of rotamers is found by an integer linear programming technique. The second stage refines the backbone of the receptor. This is an iterative procedure. In each iteration, the van der Waals forces that the ligand applies on the receptor are calculated, and the normal mode with the best correlation to these forces is identified. Then, and the receptor’s backbone conformation is minimized along this normal mode. In the next stage, the rigid-body orientation of the ligand is optimized by a Monte-Carlo (MC) technique. In the final stage, the refined models are ranked according to a binding energy function.

Experimental Results

The method was tested on a set of 20 protein-protein complexes in which the receptor’s interface RMSD, between its bound and the unbound conformation, varies in the range of 0.59Å – 4.15Å. In order to evaluate the contribution of the backbone flexibility modeling within the docking refinement process, we compared the performance of the method with and without the backbone refinement procedure. We performed three experiments on each of the 20 complexes in the data-set. In the first experiment we tested the performance of the method in refining a complex structure, in which the ligand is in its native binding geometry and the receptor is in its unbound conformation. In the second experiment we refined, for each test case, 500 randomly generated near native docking solutions. In the last experiment we refined the best 500 results of the PatchDock rigid-body docking method and re-scored the results. The results show that the method successfully models backbone movements which occur during molecular interactions. Moreover, inclusion of the backbone refinement procedure improves both the accuracy and the ranking of near native docking solution candidates.

Discussion

We present a new method for flexible refinement of docking solution candidates. The method models both side-chain and backbone flexibility and performs rigid body optimization on the ligand orientation. The refinement algorithm mimics an induced fit process. The backbone and side-chain movements are modeled according to the binding van der Waals forces between the receptor and ligand. Unlike previous methods, FiberDock uses both low and high frequency normal modes and therefore is able to model both global and local conformational changes.

Currently, the method is most helpful in cases in which the conformational changes are induced by steric clashes. In the future we plan to incorporate additional chemical forces in the backbone refinement procedure, such as electrostatic forces. In addition we plan to simultaneously model backbone flexibility of both the receptor and the ligand. This can be achieved by a minor modification in the backbone refinement procedure, by choosing the most relevant normal mode among a set of both the receptor’s and ligand’s normal modes.

In order to make the FiberDock method available for the entire biological community, a clear and user friendly web-server was developed. This is the first web-server for flexible docking refinement which models backbone flexibility and it is freely available at: http://bioinfo3d.cs.tau.ac.il/FiberDock/.


Alessandro Pandini and Jens Kleinjung, Division of Mathematical Biology, National Institute for Medical Research, UK.
We present a new approach to extract representative fragments from attractors in the space of possible local protein structures. Highly populated attractors are expected to be the most energetically stable conformations. The density based structural alphabets are optimal to reconstruct known proteins and informative to describe the most favorable conformations.

Methods:
Representative fragment selection was performed on a set of high quality structures from the ASTRAL SCOP database[3]. Structures with less than 10% sequence identity and a SPACI score greater than 0.5 were included.

Each domain was split into overlapping Calpha-only fragments of length four residues. The fragment geometry was defined by two pseudo angles and a pseudo torsional angle calculated from the Calpha positions.

The OPTICS algorithm was applied to create a Reachability Plot. In this plot, objects are ordered according to a nearest neighbor procedure and the value of their Reachability Distance (RD) is recorded. Since the RD provides a measure of connectivity, clusters of points become dents in the plot: a group of connected objects starts with a decrease in RD and ends with a new increase. Unique clusters were extracted from the Reachability Plot by a simple and robust iterative procedure and representative fragments were identified by their surrounding density.

The procedure was repeated for different density values and each structural alphabet was assessed by measuring the root mean square deviation in protein reconstruction[4] of a set of high quality structures.

Results:
We derived fragment alphabets with sizes in the range 20-40 fragments and with the ability to reconstruct known structures with an average RMSD error of 0.6 – 0.9 Angstrom on the Calpha positions. By extensive comparison with previous alphabets[1], we demonstrated that our density-derived alphabets outperform geometrically derived sets and compete with alphabets extracted by machine-learning approaches.

The OPTICS approach allowed us to recover both relatively frequent and rare fragments. The former are describing the geometry of secondary structure elements and are usually oversampled in structural databases. The latter can capture irregular loop conformations usually included in functional regions. Since the representative fragments were selected by highest density, they also define the most energetically stable conformations. Therefore each alphabet is a consistent set of attractors in the space of possible local protein structures and each representative fragment is characterized by its geometry and by the value of its surrounding density. This information can be exploited to analyze protein stability and dynamics.

References:


Here, we present our current non-redundant dataset of protein–protein interfaces and its comparison with the older datasets. We found that the number of unique interface architectures continues to grow up and the functional coverage is not complete. Some architectures are more favorable and frequently used in protein–protein associations.

The diverse range of cellular functions is performed by a limited number of protein folds and unique interface architectures existing in nature. Here, we present 8205 interface clusters, each representing a unique interface architecture. This dataset of protein-protein interfaces is analyzed and compared with older datasets to examine whether the increase in the number of unique interface architectures is leveling off and whether the functional coverage and interaction maps in the PDB are complete. We demonstrate how far we are from the complete structural information and where we are currently in structural biology. As a result, we found that the number of unique interface architectures continues to grow up. The generation of the current data set with new PDB structures illustrates both (i) an increase in the number of interface clusters, i.e., newly discovered interface architectures and (ii) an increase in the population of clusters, i.e., more interfaces with a given architecture. The current number of distinct clusters allows analysis of the functional divergence and evolution of the interfaces. If two chains interact with each other through an interface, the functions of the chains should also be related. We observe that the functional coverage is still not complete in PDB and has started to grow slowly; yet there are still unconnected functional classes indicating that this network will continue to develop in the coming years. Although the number of interactions between the functions increases continuously, there is a jump in 2002 that corresponds to the completion of several genomes or the starting outcome of structural genomics.

Some interface architectures are more favorable and frequently used in protein–protein associations. In agreement with the broadly-accepted notion that binding and folding are similar processes, we observe that most populated folds are structurally similar to the most populated interface architectures. For example, 1o1pAD is the representative interface of the most populated cluster and the three-helix bundle fold (1enh) is one of the most populated folds. 38 out of 54 residues in 1enh match structurally with the interface residues in 1o1pAD. The RMSD value is 1.86 Ångstrom (considering alpha carbon atoms) for this match, which is an appropriate value to declare that the structure of the interface 1o1pAD is similar to that of the three-helix bundle fold. In the picture, the ribbon diagrams and the matching parts of 1enh and 1o1pAD are shown. The red parts are structurally matched residues; the yellow parts are unmatched residues. The figure highlights this similarity visually. Nature appears to use similar preferred fold templates for single chains and for interfaces.

Further, using this interface dataset, we extract domain fusions in PDB. We observed a consistent trend that complex organisms prefer intra-chain domain interactions; others, less complex organisms, prefer inter chain interaction. This adaptation may be the result of the crowded traffic in the eukaryotic cells.

Recently, we also divide our dataset into three classes according to their interface similarities and global fold similarities: i. Type I clusters; same domain pairs similar interface architectures, ii. Type II clusters; similar interface architectures coming from dissimilar folds which imply that differently folded proteins can interact through similar binding sites. iii. Type III clusters; one side similar architectures coming from dissimilar folds which shows that one partner can interact with various partner chains. These subsets give information about binding specificity and selectivity of the proteins and conservation of the interaction between different protein pairs. Our structurally non-redundant dataset is a rich source for studies about protein – protein interactions; i.e. prediction of new protein interactions by structural matching (PRISM, http://prism.ccb.ku.edu.tr/prism), identification of computational hotspots in protein interfaces (HotSprint, http://prism.ccb.ku.edu.tr/hotsprint). We believe that this comparative and multidirectional analysis of the interface datasets shed light on the way of structural biology; especially, in detection of the binding region patterns, protein function prediction, motif extraction, construction of protein networks.

15. Insights into conformation change on protein docking using normal mode simulations.


Protein-protein docking involves conformational change ranging from small (< 1 Ångstrom RMSD) to far larger. We have performed coarse-grained normal mode simulations on unbound molecules. We find that these simulations suggest regions of conformational change and inform if the protein may undergo substantial conformational change (>2 Ångstrom).

MOTIVATION Proteins undergo conformational changes on complex formation with other proteins. Consequently several hypotheses have been proposed to explain the recognition process beyond the classic lock-and-key of Fisher (1). Koshold (2) proposed the induced fit model in which the structural change is the result of the molecules changing conformation when they recognise one another. Another hypothesis is conformational selection in which each molecule adopts a series of conformations and upon recognition one is selected (e.g. 3). The prediction of conformational
normal modes can suggest if a protein will undergo a substantial conformational change on docking (> 2 Angstrom RMSD). By an appropriate summation over a function of the frequencies of the normal modes, we can estimate the extent of predicted conformational change. These predicted conformational changes were then plotted along the x-axis in order of increasing predicted conformational change. Remarkably, we found that nearly all the proteins observed to have substantial conformational change on docking (>2 Angstrom RMSD) lie in the highest 40% of the x-axis. However many protein observed to have low conformational change on forming this complex also lie in this upper 40%. Thus we can suggest that a protein will NOT undergo conformational change on docking. There also is a good agreement between protein size and the extent of conformational size, but the agreement with the normal mode spectrum is significantly better.

For a protein to undergo conformational change, there must be substantial mobility in the unbound isolated molecule. This would be consistent with the recognition model of conformational selection.

IMPLICATION FOR PREDICTIVE DOCKING Normal modes could, in several proteins, be used to identify flexible regions on docking. However in general, the prediction of the direction of motion may well prove to be insufficiently accurate to be a powerful tool in adjusting the conformation of the isolated unbound molecule. Of particular importance is that the frequency plot enables one to identify if quasi main-chain rigid body docking may well prove effective in predictive docking or one should aim to model main-chain flexibility on predictive docking.


16. Sequence and Structural Features of Enzymes and their Active Sites by EC Class.

Tracey Bray, Andrew Doig and Jim Warwicker. The University of Manchester, UK.

We have undertaken a systematic study of differences in sequence and structural features of enzymes and their active sites between the six main EC classes. Many features differ significantly between EC classes, including active site non-polarity, secondary structure content, amino acid compositions and size of the enzyme structure.

The number of protein structures available in the Protein Data Bank (PDB) has increased more than five-fold over the last 10 years. A large and growing number have no functional annotation and much
effort is being invested into methods to predict protein function both via homology and ab initio methods. Enzymes of similar function, whether or not they are evolutionarily related, have been shown to exhibit shared sequence and structural characteristics [1, 2]. It has also been shown that simple protein structural features, such as secondary structure content and amino acid surface fractions, were of value in predicting the top EC class for an enzyme [3]. The complexities of the machine learning methods used, however, make it difficult to deconstruct the exact relationships between features and enzyme class. Here, we look at how simple sequence and structural features, both of the whole protein and specifically of the active site, differ over the six EC classes.

Features including amino acid compositions, secondary structure content, charge fractions, average hydrophobicity score, B-factors, average isoelectric point, and surface area, both for the total enzyme and the active site region, were calculated for a non-redundant set of enzymes for which the active site location was known. Each feature was evaluated for significant differences in the distribution of values between enzymes in the six top EC classes. Several features exhibited significant differences between functions, including active site polarity, enzyme size and active site amino acid propensities. As many of these features are not independent of each other they were reduced to a list of the most significantly-different non-correlated features. The size of the enzyme’s biological unit, the proportion of the active site that is non-polar and the active-site aspartic acid content was amongst the most significantly different non-related features and these were investigated further.

Protein Structure Size

All but one of the size-related features were found to show significant differences between the six classes, with the number of residues in the biological unit being the most significantly different. The lyases had the highest number of residues in the structure, despite not having the longest sequence lengths, due to their preference for forming higher-order oligomers over other classes.

Enzymes that form oligomers may have the advantage of using their quaternary structure as a further level of catalytic control. For the multimeric protein structures in our dataset, we assessed whether the active site of the enzyme was close to or at a subunit boundary as a means of estimating how many of the enzymes may have their action regulated by the formation of the multimeric complex. We found that the class with the highest percentage of oligomers (lyases) also had the highest percentage of oligomers with shared-subunit active sites (see Figure 1). This suggests that the over-representation of oligomers in EC4 is influenced by their inclination to form active sites near to subunit interfaces.

Enzymes that exist at highly loaded points, or choke points, in metabolic network often require high levels of control due to their metabolic importance. A study of the yeast metabolic network showed that they lyases were over-represented at these metabolically important points, suggesting a trait for oligomerisation as a means of control in lyases.

Active-site Non-polarity

The proportion of the enzyme active site that is non-polar was also found to be one of the most significantly different features over the six EC classes. The oxidoreductases had the highest proportion of active-site non-polarity. It was hypothesised that the oxidoreductases’ preference for using cofactors with non-polar regions, such as NAD and FAD, could explain the increased non-polarity of their active site. Upon removing cofactor-binding enzymes from the set there was no longer any significant difference in active site non-polarity over the six classes and the distribution of non-polar active-site content in the oxidoreductases reduced significantly.

Active-site Aspartic Acid Content

Another of the most significantly-different features was the active-site aspartic acid content, with the oxidoreductases having the lowest proportion of active-site Asp. It is expected that a negatively charged amino acid, such as aspartic acid, would be selected against in an active site that has a preference for being non-polar, such as the oxidoreductases. However, this is not mirrored in the other negatively charged residues, such as glutamic acid. Indeed there is an unusual preference for Glu over Asp in oxidoreductase active sites. We found no evidence of a link with the oxidoreductases tendency to bind cofactors. There were, however, differences in patterns of hydrogen bonding in active-site Asp and Glu between the oxidoreductases and any of the other classes. An obvious feature of oxidoreductases is that they require electron transfer, often compensated by proton movement via hydrogen bond networks. It is possible that the preference for Glu over Asp relates to an adaptation related to charge transfer, but in a complex manner that remains to be established.

This systematic study of novel differences in structural features between enzyme function sheds new light on the relationship between protein structure and function. We have shown how three of the significantly different features directly may relate to the enzyme’s catalytic action and in turn such knowledge may aid the development of further methods to predict protein function from structure without the use of alignments and in enzyme design.

Figure 1. The percentage of oligomers that have single sub-unit or shared sub-unit active sites in each class.

References:

17. Three-Dimensional Structural Determinants of Amino Acid Conservation in Proteins.

Sungsam Gong and Tom Blundell. Department of Biochemistry, University of Cambridge, UK.

We wish to answer the question “what determines amino acid substitutions in the three-dimensional structures of proteins”. We show that solvent accessibility is the most important determinant, followed by the existence of hydrogen-bonds from the sidechain to mainchain functions and the nature of the element of secondary structure.

The local environments of individual amino acid sidechains restrain the accumulation of amino acid substitutions as proteins undergo neutral evolution. One of the restraints arises from the need to maintain three-dimensional structure in order to retain function. An ESST (Environment Specific Substitution Table) describes the substitution of amino acids as a function of structural environments which restrict the possible and allowable substitutions [1, 2, 3]. The combination of environmental descriptors for solvent accessibility, secondary structure and sidechain hydrogen-bonding gives 64 matrices for each amino acid in this model and each is associated with a distinct pattern of amino acid substitution.

We will demonstrate how local environments affect the substitution pattern of amino acids. In particular, a solvent inaccessible environment restricts the possible substitution of amino acids most strongly, enhancing the diagonal of the substitution matrix, but secondary structure and the existence of sidechain hydrogen-bonds also lead to different substitution patterns. We will further demonstrate the relative importance of local environments by an analysis of distances amongst the 64 tables – each characterised by a different set of restraints – followed by Principal Component Analysis (PCA) based on a matrix of substitution profiles for all 64 environments over 441 (21*21) possible substitutions (note that we distinguish cysteine (C) with a free sulphydryl group from half-cysteine (J) which participates in a disulfide bridge).

The matrices for the 64 environments (ENVs) form three distinct clusters, one of which is characterised by of the presence of a positive-phi mainchain torsion angle, while the remaining two are distinguished by solvent accessibility. A positive-phi torsion angle can be accommodated by a glycine, which has no sidechain, but for most other L-amino acids it leads to disallowed interactions between sidechain and mainchain atoms. However, for L-amino acids such as aspartic acid or asparagine interactions between the sidechain carbonyl group with the carbonyl of the mainchain peptide bond can give rise to relative stabilisation of a conformation with a positive-phi angle [4]. Indeed in our dataset, glycine occupies 63% of total amino acids in a positive-phi torsion angle followed by asparagine (8%) and aspartic acid (5.3%). In addition, within a positive-phi class, solvent accessible amino acids occur five times more frequently than inaccessible residues whereas the ratio is 2.2 fold on average for all classes. Hence, the predominance of glycine and polar residues in both accessible and inaccessible positions makes solvent (in)accessibility a less important determinant of amino acid substitution for the positive-phi class.

Each of the three major clusters is further divided by the presence or absence of hydrogen-bonds from sidechains to mainchain amides. Hence, under either solvent accessible or inaccessible environments, the establishment of hydrogen-bonds from sidechains to mainchain amides restricts the substitution of amino acids regardless of the local secondary structure. Interestingly, secondary structure leads to clusters only within each of the clusters divided by the presence or absence of hydrogen-bonds from sidechains to mainchain amides. We further looked at the clustering pattern of hydrogen-bonds to mainchain carbonyl groups and found that they are grouped together within the cluster of the element of secondary structure, but the pattern is weaker than for other environments. Our finding suggests that the types of hydrogen-bonds make hierarchical effects on the substitution pattern of amino acids; hydrogen-bonds between sidechain to mainchain amide groups are most influential, followed by mainchain to mainchain, sidechains to mainchain carbonyl groups. We further investigated this pattern by averaging the effect of the solvent accessibility and then both solvent accessibility and the type of secondary structures; we confirmed that the clustering and PCA results hold the same order of hierarchy. Especially, it is very evident that the influence of the eight types of hydrogen bonds from sidechains to mainchains is very hierarchical.
It has long been understood that hydrogen bonds play a very important role in the stability of a protein structure, and provide restraints on the substitutions of amino acids during evolution by neutral drift. Recently, Worth et al. addressed the importance of hydrogen-bond potentials from sidechains in the stability of protein structures [5]. They showed that the formation of hydrogen bonds to mainchain amide atoms influences conservation of amino acids, with those satisfied buried polar residues that form two hydrogen bonds to mainchain amides being significantly more conserved than those that form only one or none. Their evidence and our findings provide new insights into the roles of networks of hydrogen bonds within the three-dimensional structures of proteins.

References

18. Structural rearrangement in the TCRpMHC formation in reaction to agonistic and antagonistic peptides.

Bernhard Knapp and Wolfgang Schreiner. Medical University of Vienna, Austria.

We present series of molecular dynamics simulations of two different TCRpMHC complexes. The first complex contains an agonistic peptide, the second an antagonistic one. The structural rearrangement in the CDRs of the TCRs in reaction to the different peptides is characterized.

The interaction between the T-cell receptor (TCR) and the major histocompatibility complex (MHC) is essential in immunology. The antigen presenting cell (APC) cleaves proteins into small peptide (p) fragments and transports them to the cell surface. These epitopes are presented by MHC molecules to cells of the immune system and recognized by the TCR. The hypervariable complementarity determining regions (CDRs) play a key role in this interaction between the MHC, the peptide and the TCR.

Many “in silico” methods for the prediction of the binding affinity between the MHC and the peptide are available but only few studies investigate the deformation of the CDRs in reaction to different peptides. In this study we show the structural rearrangement in the TCR HA1.7 in reaction to an agonistic and an antagonistic peptide presented by the DRB1 MHC-allele. Series of molecular dynamics (MD) simulations starting from different initial forces were carried out to investigate the deformation of the CDRs: The TCRpMHC molecules were immersed into an artificial water bath and energetically minimized. Thereafter, the system was warmed up to 310 K and the MD simulations were carried out for a real time of 20 ns.

Interesting differences were observed between the agonistic and antagonistic complexes. However, it is obvious to the authors that more complexes will be tested in future work to validate the results and to deduce general rules for the structural rearrangement of MHC and TCR molecules in reaction to agonistic and antagonistic peptides.


We consider a recently suggested “equation of state” for natively folded proteins that is based on a fractal viewpoint of proteins and on a generalization of the Landau-Peierls instability. Using the Gaussian network model, we verify its validity for about 5800 proteins.

It has recently become evident that natively folded proteins exhibit a fractal character that exemplifies in a few, well known, broken dimensions (Burioni et al., 2004; Elber and Karplus, 1986; Enright and Leitner, 2005; Granek and Klafter, 2005; Lushnikov et al., 2005; Reuveni et al., 2008; Stapleton et al., 1980). The mass fractal dimension describes the spatial distribution of the mass within the protein (Stauffer and Aharony, 1994) and the spectral dimension governs the density of low frequency vibrational normal modes.

Daniel Rigden, University of Liverpool, UK.

After heroic efforts by bioinformaticians, accurate ab initio modelling is now possible for many small proteins. Results from three distinct areas of application are presented here. Structure-based function predictions are given, as well as novel applications to post-process database search results and to solve crystal structures by Molecular Replacement.

As recently as 1997, reporting on CASP2, Arthur Lesk found ab initio modelling results “disappointing, even sobering” and related that, with a single exception, they showed no more than “fragmentary success” [1]. Nevertheless, while size restrictions remain, the past few years have seen dramatic developments transforming ab initio modelling of protein structures into a practical functional reality [2]. Significantly, some algorithms have been publicly released, enabling third parties to assist in methodological improvements, and others are available as servers. Given this availability the time is ripe to encourage wider adoption of ab initio modelling and to consider possibilities for its integration with other areas of protein bioinformatics and biosciences in general. This work reports several promising directions in which ab initio modelling can be used to explore protein structure, function and evolution. In particular, novel applications in post-processing borderline database search results and interfacing with experimental structural biology are presented.

Perhaps the best established application for ab initio modelling is for structure-based function prediction. It can be used as a valuable reinforcement for borderline fold recognition results, but also effective alone. This work describes several applications of this approach such as assignment of a role in control of gene expression for the bacterial sporulation protein SpoVS [3]. Recent results on ab initio modelling of ORFans and viral proteins are also presented. In
both cases the absence or paucity of sequence relatives makes it harder to locate distant homologues, even by sensitive profile-profile or fold recognition methods. Nevertheless, ab initio modelling can step in where comparative modelling is impossible. In particular, it is shown that nucleic-acid binding proteins may often be identified using methods that can spot their shared characteristics. The appearance of familiar folds in ab initio models often sheds light on protein evolution. For example, models of Stx (Shiga toxin) phage-encoded proteins, supported by genomic context and other predictions, suggest roles in DNA replication or transcription. One protein, lying close by the phage’s helicase, resembles the inserted domain found in many helicases but not, suggestively, the phage’s own, hinting that it may resemble an ancestral protein acquired by the helicases in some lineages. It is also intriguing to note that several ORFans found in the intracellular parasite Anaplasma phagocytophilum are predicted to most closely resemble eukaryotic proteins suggesting that they may have been acquired from the host.

The first novel application here presented is the sifting of borderline database search hits. Commonly, clear hits are not well separated from evidently unreliable matches: there is often a grey area in which sequence statistics alone are incapable of inferring (in)significance. Such situations are common for small sequences, which contain less information than larger ones, and for highly divergent families. Using the MIT (microtubule-interacting and trafficking) domain as an example, the first use of ab initio modelling for mining the borderline significant zone of database search results is presented [4]. It is shown that sequences with e-values as high as 0.9, against a maximally sensitive structure-informed PSI-BLAST profile, in fact contain genuine MIT domains. Importantly, non-homologous sequences scoring e-values as low as 0.07 are unambiguously distinguished from MIT domains by the ab initio modelling.

The solution of crystal structures by Molecular Replacement (MR) using ab initio models is also presented. Following a single proof of principle that employed large-scale distributed computing resources [5] it is shown, for a set of proteins, that models produced with typical hardware can be sufficiently accurate [6]. Results with a set of 16 trial proteins showed that models, sometimes with limited manual clipping, either singly or in combination, could solve five crystal structures. There is an advantageous correspondence between the clustered models produced by ab initio procedures, found to more accurate in the core fold, and the ability of MR to work with ensembles of models and/or incomplete models. The five successes were each obtained when ab initio modelling produced large top clusters of models showing that the likelihood of MR success can be predicted. Avenues to improve and automate the methodology are suggested. Finally, recent data are presented that show that inserted domains can be effectively modelled ab initio and the results successfully used for MR.

Ab initio modelling is still limited to sequences of 100-120 residues and is not always successful, yet it is clear that it already has much to offer study of protein structure, function and evolution. Continued methodological progress and ever-increasing computer power will clearly extend the limits of ab initio modelling in the future. The possibilities for its application in the areas above, and no doubt others, can only continue to grow.

References


21. Genome sequencing delivers diminishing returns for homology detection.

Daniel Chubb, Lawrence Kelley and Michael Sternberg, Imperial College London, UK.

It is a commonly held belief that genome sequencing aids in the identification of remote homology between proteins by bridging gaps in sequence space, thereby expanding functional and structural annotation. We challenge this belief by rolling back the sequence databases of the last two decades and observe what effect the sequence explosion has had on our ability to identify remote homology. We have discovered that using conventional methods we are already facing diminishing returns and therefore must continue to develop novel methodologies to reap the maximum benefit from new sequence data.
Assigning family membership(1), predicting function(2), locating binding sites(3), and predicting structure(4, 5) all rely upon the accurate identification of homologs whose features have been experimentally determined.

The degree of similarity between protein sequences tends to be greater for homologs than for unrelated proteins. Accordingly, most homology identification methods rely upon the accurate alignment of sequences and the quantification of this sequence similarity. This is a trivial task for closely related proteins with high sequence ID, but as relationships become more distant, sequence similarity diminishes and becomes increasingly hard to detect. The most sophisticated methods iteratively build up profiles of sequences, starting with confident homologs. These profiles record the propensity for each amino acid (and sometimes insertions and deletions) to be present at any position and are compared against a sequence or profile library. It is widely assumed that as the sequence databases increase in size, the ability to detect remote homologs will increase as new intermediate sequences are incorporated within profiles. Here we test this notion.

We recreated the sequence databases of the past two decades by extracting sequences from the universal protein resource database (Uniprot) (6) according to their creation date, and investigated our ability to detect remote homology using sequence profiles built from each of these databases. The construction of profiles and homology detection was performed by PSIBLAST(7), arguably the tool most widely used for this task with 27,068 citations in the literature (as of January 2009). Our gold standard for homology between two proteins was membership of the same superfamily within the SCOP (8) database of experimentally determined protein domain structures. To focus on remote homology, we used SCOPO30 (9), a subset of SCOP containing sequences possessing a similarity of no more than 30% to each other. PSIBLAST profiles were created for each of these sequences by searching against each database from 1987 to 2007. Each sequence profile was then searched against the fixed SCOPO30 sequence database and the number of positive hits was recorded (from 220124 possible hits). The results are shown in the figure together with the size of the sequence database in each year.

It is clear that sequence database growth has resulted in an overall improvement in remote homology detection using PSIBLAST. However, this improvement does not directly scale with the massive increase in available sequences particularly evident in the last decade. Indeed, the increase in homology detection plateaus in 2004 and subsequently declines (from 53927 to 52397 hits in 2007). This trend remained even when the order of sequence discovery was randomised and under a range of PSIBLAST parameters. Further investigation in-to the annual change in detection showed that in every year, homologs previously identified are lost – in the case of the post 2004 results, this loss is greater than the gain of new hits. Similar results were obtained when using a more sensitive HMM-HMM comparison method: HHsearch (10).

We hypothesise that the increasing sequence coverage is not expanding our detection of homologs; it is shifting it in different directions which are reliant on the sequence composition in any given year. These shifts are likely to be influenced by – but not restricted to – a combination of sequence redundancy and high diversity within protein families. Sequence redundancy has been shown to saturate profiles with highly homologous sequences, which warp the homology detection space in their direction (11). At the same time, profiles diverse homologous families could lead to profiles that are too divergent to be able to detect all members.

The influx of protein sequences shows no sign of slowing; on the contrary, large scale metagenomic sequencing projects combined with the ever-falling cost of sequencing means that we are likely to see an acceleration in sequencing. Our results suggest that not only will this extra sequence data not be beneficial for homology detection; it could be actively harmful, adversely affecting our ability to predict structure and identify the function of novel proteins.
Depending on the accuracy of the structural information, one can reach different levels of elucidation of the relationships between structure and function. At high level of resolution, biochemical reaction may be predicted using molecular modeling and substrate docking methods [Hermann et al., 2006] [Song et al., 2007] [Hermann et al., 2007] [Kalyanaraman et al., 2008] [Xiang et al., 2009]

At low resolution level, functional annotation can be supported by fold assignment and 3D motif searching. For example, conserved structural cavities in a protein family are an indicator of active-site [Dundas et al., 2006]. Residues in these cavities are subject to different selective pressures so that multiple alignments can reveal conserved profiles. Hidden Markov Models (HMMs) provide a coherent statistical theory for this analysis.

We have developed a methodology for analysis of residues of protein cavities to detect the determinant ones that can be responsible for the protein function. Instead of using a global MSA, we use structural alignments of the predicted cavity residues. From these alignments, we are able to divide the protein families into groups of similar profiles using conceptual clustering [Fisher, 1987]. The analysis detects intra-family variations that can be responsible for function and / or specificity.

In this work, we will present our methodology that led us to propose hypothetical binding site residues for the X protein (confidential). Molecular docking was performed with the ligands and showed plausible poses for the molecules in the active site pocket. Predictions are confirmed by structural resolution and are being tested by directed mutagenesis experiment.

References


23. Effective atomic interactions for the characterization of protein-ligand binding interfaces.

Alex Slater, Evandro Ferrada and Francisco Melo. Pontificia Universidad Catolica de Chile, Chile.

We describe a new methodology for the analysis of intermolecular interfaces, which relies on the calculation of the effective atomic interactions between two molecules (i.e., those pairwise atomic interactions not occluded by other atoms). We have used this methodology to classify all protein-ATP complexes and found several examples that demonstrate its usefulness.
complex and also suggest that performing an independent analysis of each binding partner can be inaccurate to classify and characterize the known protein-ligand complexes. We emphasize that the proper classification of protein-ligand binding interfaces is important for the accurate definition of non-redundant protein-ligand binding modes and thus relevant for the development of any structure-based method for ligand binding site prediction in proteins.


Tomas Norambuena and Francisco Melo, Pontificia Universidad Catolica de Chile, Chile.

We present a new database with all protein structures from the Protein Data Bank (PDB) complexed with DNA and solved at a resolution higher than 2.5Å. Each of the 792 entries has been manually curated and annotated by humans. The database contains many annotated features and links that add value over the PDB itself and other databases.

The current version of the database contains 792 protein-DNA complexes from the PDB that were solved by X-ray crystallography at a resolution higher than 2.5Å. The database is fully linked to other servers and databases such as the PDB, PubMed, COPS and NCBI-BLAST. The database includes a function/structure-based classification of entries with three hierarchical levels. Each entry also includes categorizations resulting from the visual inspection of the respective complex. These categories concern mostly how DNA and protein are arranged in the complex. Other fields of this database have been calculated directly from the PDB atomic coordinates and include the effective atomic interfaces of each complex, the atomic coordinates of the asymmetric unit and biological molecules. The function/structure-based classification of the complex consists of four major categories: Enzyme (if the main function of the protein is to modify DNA), Transcription Factor (if the main function of the protein is to regulate transcription and gene expression), Structural/DNA Binding Protein (if the main function of the protein is to support DNA structure, bend DNA or aggregate other proteins), and Immunological Protein (if the DNA/protein interaction triggers response of the immune system). The function/structure-based classification of the complex contains several types for each category mentioned above and relies on information gathered from the PDB and PubMed databases. There are 19 types for the Enzyme category (Methyltransferase, Repair Protein, Topoisomerase, Nuclease, Recombinase, Glucosyltransferase, Transposase, Phosphodiesterase, Excisionase, Kinase, Photolyase, Helicase, Ligase, Translocase, Helicase, Endonuclease, Polymerase, Glycosylase); 8 types for the Transcription Factor category (Zinc Coordinating, Zipper Type, Alpha/Beta, Alpha Helix, Ribbon/Helix/Helix, Beta Sheet, Helix Turn Helix); 8 types for the Structural/DNA Binding Protein category (Replication, Maintenance/Protection, Recombination, Zalpha, Centromeric Protein, Telomeric Protein, Structural Protein); and only one type for Immunological Protein category (Immunoglobulin). Finally, some categories and types also have some defined subtypes that take into account structural domains, specific reactions of enzymes, specific DNA binding sites, etc. Another important aspect of this database is that it contains several fields that describe structural features of the DNA bound to the protein. These features include the annotation of the existence of single or double strand DNA, sticky ends, flipped bases, nicked DNA, gapped DNA, open DNA, modified DNA, hollyday junctions and cruciform DNA motifs. We hope that the public release of this database will facilitate the study of the key structural determinants in the specificity of DNA recognition by proteins.

25. Prediction of Functional Residues in Protein Structures with Insights into Protein Phosphorylation.


We introduce a novel graph-based kernel method for annotating functional residues in protein structures. The approach captures residue interconnectedness in local structural neighborhoods and has improved performance compared to alternative approaches, as demonstrated on the prediction of catalytic residues and protein phosphorylation sites. New insights into structured phosphosites are provided.

With over 50,000 structures deposited in the Protein Data Bank (PDB) and high-throughput efforts under way, functional characterization of proteins with known 3D structure is gaining importance in the global effort to understand structure-to-function determinants. Experimental assays for functional characterization are expensive and time-consuming, thus the development of accurate computational approaches for function prediction is
essential to the functional annotation process. Typically, the problem of protein function prediction reduces to one or more of the following questions: (1) prediction of the molecular and biological function of the molecule, (2) prediction of ligands, cofactors, or macromolecular interaction partners, and (3) prediction of the residues involved in or essential for function, e.g., interface sites, hot spots, metal binding sites, catalytic sites or post-translationally modified residues. At a higher level, computational methods can be used to establish connections between proteins and disease, typically via simulating protein folding pathways or by using statistical inference techniques to predict gene-disease associations or the effects of mutations. Prediction of protein function from 3D structure emerged in the late 1980s when the accumulation of solved structures in PDB made systematic studies feasible. There are four basic approaches used in this field, starting from residue-level function and building toward higher level annotation: (1) residue microenvironment-based methods, (2) template-based methods, (3) docking-based methods, and (4) graph-theoretic approaches. In addition to these bottom-up strategies, another group of methods tackle the problem top-down to directly predict protein function on a whole-molecule level, without necessarily finding functional residues, and then investigate the residues most critical in the classification process. In this study, we present a new method, referred to as the graphlet kernel, for identifying functional sites in protein structures. We first represent a protein structure as a contact graph where nodes are residues and edges connect vertices that correspond to the neighboring residues in space. The method then combines the graphlet representation of every vertex in a graph (Pržulj, et al., 2004) and kernel-based statistical inference. We extend the concept of graphlets to labeled graphlets and use the counts of labeled graphlets to compute a kernel function as a measure of similarity between the vertices. We show that the graphlet kernel generalizes some previous methods such as FEATURE (Bagley and Altman, 1995) and can also be readily extended to other problems involving graphs, in either a supervised or an unsupervised learning scenario. The problem addressed here can be generally defined as follows: given a protein structure, probabilistically assign function to each amino acid. Functional assignment is based on similarities of the structural neighborhoods of residues under consideration and measured in terms of local patterns of inter-residue connectivity. We start by modeling a protein structure as a protein contact graph, where each amino acid is represented by a vertex in the graph and two vertices are connected by an undirected edge if the corresponding amino acids are closer than some predetermined distance (Figure 1). Each vertex in the protein contact graph is then represented as a vector of counts of labeled non-isomorphic subgraphs (called graphlets), centered on the vertex of interest. A similarity measure between two vertices (kernel function) is expressed as the inner product of their respective count vectors and is used in a supervised learning framework to classify protein residues. Figure 1. (A) Protein structure of human lymphocyte kinase Lck with a phosphorylation site Tyr394; (B) protein contact graph for Tyr 394 with its level-3 structural neighborhood. We evaluated our method on two function prediction problems: identification of catalytic residues in proteins, which is a well-studied problem suitable for benchmarking, and a much less explored problem of predicting phosphorylation sites in protein structures. We compared the graphlet kernel approach against two alternative methods, a sequence-based predictor and our implementation of the FEATURE framework. On both function prediction tasks the graphlet kernel performed favorably compared to the alternatives; however, the margin of difference was considerably higher on the problem of phosphorylation site prediction. While there is both computational and experimental evidence that phosphorylation sites are preferentially positioned in intrinsically disordered regions, we provide evidence that for the sites that are located in structured regions, neither the information related to surface accessibility alone nor the averaged measures calculated from the residue microenvironments utilized by FEATURE were sufficient to achieve high prediction accuracy. It was previously proposed that phosphorylation sites preferentially, although not exclusively, appear in intrinsically disordered protein regions (Iakoucheva, et al., 2004). A recent mass spectrometry study of 162 cytosolic phosphoproteins provided an experimental confirmation: out of 512 phosphorylation sites, 97% occurred outside of structured domains, and 86% occurred in regions of protein disorder (Collins, et al., 2008). Nonetheless, there are examples of ordered phosphorylation sites and a number of structures containing phosphorylated sites have been deposited in PDB. After analyzing structured phosphorylation sites we conclude that (1) protein phosphorylation sites are indeed preferentially located in disordered regions; (2) ordered phosphorylation sites are preferentially located in protein loops thereby potentially facilitating the access of kinases to the phosphorylatable residue; (3) for the cases of ordered phosphorylation sites currently present in PDB there are minimal structural changes that occur upon phosphorylation with only a few examples of order-to-disorder transitions; (4) ordered phosphorylation sites, especially for threonine, are enriched among kinases; and (5) ordered tyrosine phosphorylation sites are frequently found to be autophosphorylation sites. Bagley, S.C. and Altman, R.B. (1995) Protein Sci, 4, 622-635. Collins, M.O., Yu, L., et al. (2008) Mol Cell Proteomics, 7, 1331-1348. Iakoucheva, L.M., Radivojac, P., et al. (2004) Nucleic Acids Res, 32, 1037-1049. Pržulj, N., Corneil, D.G. and Jurisica, I. (2004) Bioinformatics, 20, 3508-3515.


We present PoreWalker, a fully-automated method to detect and fully characterise transmembrane channels from their 3D structures. PoreWalker exploits a geometry-based stepwise approach and was applied to several transmembrane channel structures, effectively identifying specific functional and/or structural channel features. The software is implemented as a web-based resource at http://www.ebi.ac.uk/thornton-srv/software/PoreWalker/.

Transmembrane channel proteins play pivotal roles in maintaining the homeostasis and responsiveness of cells and the cross-membrane electrochemical gradient by mediating the transport of ions and molecules through biological membranes [1]. Understanding the structure and function of transmembrane channels and studying their properties and biochemical mechanisms is therefore a very important task in biological and pharmaceutical research [2, 3]. Transmembrane channel proteins usually show a cavity spanning the whole protein, the pore, which constitutes the path used by ions
and/or molecules to cross the membrane. The pore has two openings, one on each side of the membrane, and it has been hypothesized (and in some cases shown) that the specificity and selectivity to different solutes is strongly dependent on the particular structural or amino acid composition features of the channel [4-6]. Subsequently, computational methods for the identification and characterization of pores in transmembrane protein 3D-structures represent key tools to gain insights into how these proteins function.

To our knowledge, methods currently available for structural analysis and visualisation of transmembrane channels include HOLE, developed in 1993 and still widely used [7, 8], CAVER [9] and its improved version MOLE [10], and MolAxis [11]. The four programs are based on very different approaches, but they all require user-defined starting points and/or vectors that assume a fairly good knowledge of the location of the pore and/or of key residues lining pore walls. Moreover, they provide only a limited characterisation of the channel mainly consisting of diameter values and some of the residues lining the pore walls.

We developed PoreWalker, a novel web-available method for the detection and full characterisation of transmembrane protein channels from their 3D-structure. PoreWalker is fully automated and very user-friendly, requiring as input only the 3D coordinates of a transmembrane protein structure and uses a stepwise approach. First, the pore centre is modelled and optimised, and then the biggest and longest cavity through the channel is detected and the optimal pore axis generated. Finally, a number of pore features are calculated and a complete characterization of the channel is displayed. In particular, in addition to diameter profiles, PoreWalker characterises several specific pore features: shape and regularity of the channel cavity, atoms and corresponding amino acids lining the pore wall, and positions of pore centres along the channel. These features can be very helpful to gain insights into the functional and structural properties of transmembrane protein channels by triggering specific in silico or experimental analyses [12].

The method was applied to several structures of transmembrane protein channels and it proved able to identify shape/size/residue features representative of specific channel families, to detect important structural features such as filters, gates and cavities, and to trigger the study and understanding of functional mechanisms. In summary, PoreWalker provides a reliable and automated resource to gather data and information useful for a deeper understanding and classification of membrane protein structures. The software is implemented as a web-based resource at http://www.ebi.ac.uk/thornton-srv/software/PoreWalker/.

27. The effect of loops on the structural organization of α-helical membrane proteins.


Analysis of the x-ray structures of 41 α-helical membrane proteins suggests that interhelical loops affect the structural organization of membrane proteins. Stretched loops constrain the corresponding helices. Hydrophobic residues in other loops are preferably buried, leading the loops to collapse, which forces the TM helices to assemble.
interactions. The effect of interhelical loops on the structural organization and stability of membrane proteins has been recognized in general but specific research has been limited.

We have studied this effect of loops by analyzing the x-ray structures of 41 alpha-helical membrane proteins. First we define the loop flexibility ratio, R, and find that 33% of the loops are stretched, where a stretched loop constrains the distance between the two connected helices. The significance of this constraining effect is supported by experiments carried out with bacteriorhodopsin and rhodopsin where cutting or eliminating their (predominately stretched) loops has led to a decrease in protein stability, and for rhodopsin in most cases also to the destruction of the structure. We show that for non-stretched loops in the extramembranous regions the fraction of hydrophobic (H) residues is comparable to that of soluble proteins; furthermore (as for soluble proteins), the H residues in these regions are preferred to be buried. This is expected to lead to the compact structural organization of the loops, which is transferred to the TM helices causing them to assemble. We argue that a soluble protein complexed with a membrane protein similarly promotes compactness; other properties of such complexes are also studied. We calculate complementary attractive interactions between helices, including hydrogen bonds (HBs) and van der Waals interactions of sequential motifs, such as GXXXG.

The above interactions are divided into two groups, long- and short-range. Thus, hydrogen bonds between successive helices along the chain (as well as geometrical constraints imposed by stretched loops) are of a short-range type because they cannot lead to the assembly of the helical bundle. This assembly is created by long-range HBs and van der Waals interactions, and by the collapse of the non-stretched loops due to hydrophobicity; therefore, the long-range interactions are more important than the short-range ones. The relative and combined effects of all these factors on the association of the TM helices are discussed and protein structures with only few of these factors are analyzed. Our study emphasizes the need for classifying membrane proteins into groups according to structural organization. This classification should be considered the need for classifying membrane proteins into groups according to interactions. The effect of interhelical loops on the structural organization and stability of membrane proteins has been recognized in general but specific research has been limited.

Membrane proteins are essential parts of all living cells. Due to their importance in medically relevant processes, membrane proteins constitute the majority of current drug targets. Knowing the structures of membrane proteins is thus of great value. However, experimentally determining the structure of membrane proteins is difficult and computational methods have so far only focused on modelling soluble proteins. Homology modelling is currently the most successful method of predicting protein structure. Its major requirement is the knowledge of the structure of a homologous protein. Comparable to when the first homology modelling methods for soluble proteins were developed, there are now enough known membrane protein structures to make this approach applicable to membrane proteins. We have taken the first steps towards creating such a method. The hydrophobic core of a biological membrane constitutes a radically different chemical environment to the aqueous solvent in which globular proteins reside. It is thus important to know which parts of the protein reside within the membrane, and which parts are not. iMembrane (http://iemembrane.info) is a novel method that attempts to solve this problem. It takes as input a membrane protein structure (or sequence) and annotates every residue with regard to its position within the membrane. This is done by using the principle of homology to exploit existing data of proteins within membranes. The data is simplified and projected onto the input protein, making it applicable to a general modelling framework. For each residue in the target protein, two types of annotation are generated: whether or not it is in direct contact with the membrane, and its position in space relative to the membrane (see attached figure, showing the latter annotation type. Red = "residues in the hydrophobic membrane core layer", white = "residues in the polar head group layer", dark blue = "residues outside of the membrane"). iMembrane has enabled the creation of substitution tables specific to the various membrane environments. These tables showed notable differences in substitution behaviour between membrane and non-membrane environments, as well as differences between the different environments within the membrane. We are also working on using iMembrane to create model quality assessment programmes (MQAPs) for membrane proteins. These and future projects will lead to the eventual development of a framework for the complete structural modelling of membrane proteins.

For more information about iMembrane, please refer to: iMembrane: Homology-Based Membrane-Insertion of Proteins Sebastian Kelm; Jiye Shi; Charlotte M. Deane Bioinformatics 2009; doi: 10.1093/bioinformatics/btp102

Stefano Lise, Cedric Archambeau, Massimiliano Pontil and David Jones. University College London, Dept. of Computer Science, UK.

Protein-protein interactions are critically dependent on just a few residues (hot spots), which if mutated can disrupt complex formation. We present a novel computational approach to identify hot spot residues, given the structure of a complex. The approach combines the strengths of machine learning and energy-based methods.

Protein-protein interactions (PPI) are central to most biological processes including for example cellular communication, gene regulation, and immune response. The complexity of these processes, coupled with the intricate interaction network that biomolecules form in a cell, requires proteins to be able to selectively bind to other proteins. Indeed, erroneous or disrupted protein interactions can be the causes of a number of diseases. Elucidating the fundamental biophysical principles that govern molecular recognition and drive protein association is therefore a topic of primary importance in biomedical research. Current knowledge about PPI has been derived from structural, thermodynamic and mutational studies. However, despite many progresses, a comprehensive understanding of the physical basis of affinity and specificity in PPI is at present still lacking and fundamental problems related to the recognition process are yet to be solved. Alanine scanning mutagenesis is a powerful experimental methodology for investigating the structural and energetic characteristics of protein complexes. Individual amino-acids are systematically mutated to alanine and the effects on binding affinity observed. As alanine amino acids do not have a side-chain beyond the beta-carbon, this procedure in effect determines the relative importance of each side-chain to complex formation, providing a map of the so-called functional epitope (to be distinguished from the structural epitope defined instead by all residues at the interface). Several experiments have shown that protein-protein interactions are critically dependent on just a few residues (hot spots) at the interface. Hot spots make a dominant contribution to the free energy of binding and if mutated they can disrupt the interaction. For the majority of interface residues instead, the effect of an alanine mutation is minimal.

As mutagenesis studies require significant experimental efforts, in recent years a number of computational approaches have been developed to predict hot spot residues in a protein complex structure. Most methods consider an energy function $G$, model an alanine substitution and estimate the induced change in binding free energy ($\Delta G$), e.g. by means of molecular dynamics simulations or of empirically calibrated energy functions. More recently, machine learning approaches have also been applied to the problem of detecting hot spot residues. Accurate predictive models provide a valuable complement to experimental studies and add to our understanding of the factors that influence affinity and specificity in protein-protein interfaces. In addition, they can have important applications in the field of drug discovery as PPI are emerging as a new class of potential targets for therapeutic intervention. Although dealing with protein binding epitopes is more challenging compared to, e.g., enzyme binding pockets, a number of studies have been successful in developing (drug-like) small molecules that bind at hot spots and inhibit complex formation. Reliable hot spots predictions can therefore represent the first step in rational drug design projects.

Here, we present a novel computational strategy to identify hot spot residues, given the structure of a complex. Similarly to other energy-based methods, we consider the basic terms that contribute to hot spot interactions, i.e. van der Waals potentials, hydrogen bonding, electrostatic interactions and solvation energies. Rather than writing an explicit energy function from which we can then calculate $\Delta G$, we treat them as input features of a machine learning algorithm. The rationale beyond our approach is that the exact functional form for $\Delta G$ is not known but it is reasonable to assume that it would incorporate these terms. Given a set of training examples of alanine mutation data, we use machine learning methods to optimally combine and integrate them and to deduce the functional properties of $\Delta G$.

We show that our approach provides significant improvements over available methods in predicting hot spots (here defined as those residues for which $\Delta G > 2$ kcal/mol). In particular we find the best performances using Transductive Support Vector Machines, a semi-supervised learning scheme. We point out shortcomings and limitations of our method and suggest possible strategies to improve it. The hybrid scheme we have developed combines the strengths of machine learning and energy-based methods. Although so far these two types of approaches have mainly been applied separately to biomolecular problems, the results of our investigation indicate that there are substantial benefits to be gained by their integration. We discuss potential applications to other problems of interest.

30. Assisted crystallographic RNA model
building: A directed rotameric approach for building the RNA backbone.


The backbone of RNA is critical for function, but studies of the backbone have long been hampered by the difficulty of accurately determining its structure. We have combined a reduced representation of RNA with an all-atom rotamer library to increase the ease and accuracy of crystallographic backbone structure determination.

Recent structural and biochemical studies have shown the importance of the RNA backbone in both catalysis and interactions with other biomolecules. However, extensive studies of the backbone have long been hampered by the difficulty of accurately determining its structure. The large number of torsional angles per nucleotide makes study difficult for both X-ray crystallography and NMR. The challenges of RNA crystallography are further exacerbated by the difficulty of obtaining well-diffracting crystals, resulting in resolutions much lower than what is commonly seen in protein crystallography.

Additionally, when building protein structures into electron density, there exist many tools to aid the crystallographer in the building process, including fully- and semi-automated techniques for both building and correcting structures. However, tools for RNA structural analysis are only beginning to emerge [1-4], and there are currently no tools or methodologies for accurately building RNA backbone structure.

A previously developed RNA pseudotorsional system [1-2] allows for accurate and automated analysis of low resolution structures. This system involves two virtual dihedrals that use only the phosphate and C4' atoms: eta (C4'<n-1>, P<n>, C4'<n>, P<n+1>) and theta (P<n>, C4'<n>, P<n+1>, C4'<n+1>). These pseudotorsions can be plotted in two dimensions, similar to a Ramachandran plot, which allows for simple categorization of the RNA backbone.

However, this analysis does not alleviate the difficulty of accurately locating backbone atoms in low resolution structures. To address this, we have combined the pseudotorsions with the recently published all-atom consensus backbone rotamer library [5]. This rotamer library divides the backbone into suites, which span two sugars and the intervening phosphate, as opposed to the more traditional nucleotide division of the backbone, which spans two phosphates and the intervening sugar. By combing the pseudotorsions and this rotamer library, we have developed a methodology to automatically and accurately construct the RNA backbone starting from only the phosphate and base locations.

This methodology uses a modified pseudotorsion system that involves the C1' atom in place of the C4' atom, as the C1' is covalently bound to the base and therefore can be more reliably located in electron density. These modified pseudotorsions are used to predict the probability of each rotamer for each suite, with predictions being calculated using a Gaussian clustering [6] of rotamers in theta-eta space. These probabilities are then combined with predictions of the sugar pucker, based on the measurement of the perpendicular distance from the glycosidic bond to the 3' phosphate [3], as well as additional probabilities based on the inter-C1' and inter-phosphate distances. The highest likelihood sequence of rotamers for the entire structure is then calculated using a Hidden Markov Model, and these rotamers are used to accurately build the entire backbone structure.

This methodology can easily be incorporated into the structure building process. We are building a Coot plugin to assist the crystallographer in placing phosphate and base atoms into the electron density map. Coordinates for backbone atoms are then automatically calculated and constructed, and the crystallographer is able to modify any rotamers that do not provide a satisfactory fit to the electron density. This process will allow for more automated and more accurate determination of the RNA backbone in low and intermediate resolution structures.


A coarse-grained model based on pairwise interaction energies of amino acid residues was developed to predict stable regions in membrane proteins. The approach was applied to bacteriorhodopsin wild type and several mutants. The prediction results were compared with all-atom molecular dynamics (MD) simulations and single molecule force spectroscopy (SMFS) experiments.

Bacteriorhodopsin is a membrane-bound energy-transducing enzyme. It uses light energy to transport protons across the bacterial membrane against the proton electrochemical gradient [1]. Bacteriorhodopsin belongs to a special class of archaeal transmembrane proteins and consists of seven closely packed alpha-helices (A to G). The seven helices are arranged in two arcs, which form a membrane spanning cavity. A retinal cofactor, which is bound to a lysine residue in the middle of helix G, lies in this cavity. The mechanism of proton transport by bacteriorhodopsin is based on the light-activated isomerization of the bound retinal cofactor. Bacteriorhodopsin has been a subject of intensive experimental and computational studies, but still the folding process and the intraprotein interactions of bacteriorhodopsin are not fully understood. Misfolding of membrane proteins plays an important role in a number of diseases. Understanding of the forces contributing to the stability of bacteriorhodopsin would give insights to intraprotein interactions in other alpha-helical membrane proteins.

We have developed a coarse-grained model based on pairwise interaction energies of amino acid residues to address the problem of intraprotein interactions in membrane proteins. The effects of mutations in the stabilizing regions can also be predicted. We use this model to predict stable regions in wild type and mutant bacteriorhodopsin. The computational predictions were done on four bacteriorhodopsin mutants (P50A, M56A, P91A, P186A) for which there are experimental data available. Predictions were evaluated against experimental data from single molecule force spectroscopy (SMFS) experiments [2] and computational results from all-atom molecular dynamics (MD) simulations. SMFS is a novel technique, which measures the forces necessary to pull a protein out of the native membrane. Such force data contain valuable information on inter- and intra-molecular interactions that stabilize the protein. The evaluation against SMFS data shows that the coarse-grained model predicts all main unfolding barriers at 100% and all observed barriers including minor ones which are not present in all experiments at 61% accuracy in wild type bacteriorhodopsin. In molecular dynamics simulation trajectories 89% of the experimental barriers were detected. Hence, our coarse-grained model gives almost the same accuracy as the detailed all-atom model, but with much less computational cost. This method is the first coarse-grained model, which can predict stabilizing regions and residues in membrane proteins.


An approach to model three-dimensional structures of large protein complexes by combining various computational methods and high-quality experimental data was developed. An ensemble of structural models of histone methyltransferase complex from Saccharomyces cerevisiae was produced using this technique.
of several experimental techniques [1]. Recently there have been several attempts to produce computationally either atomic models or more coarse grained architectural models of large protein complexes. Aloy et al. started with a set of yeast protein complexes identified by tandem affinity purification and selected the most promising ones for electron microscopy [2]. They modelled the interactions between the component proteins using their similarity (sequence or structural) to interacting proteins of known structure and the obtained electron microscopy maps. Alber and coworkers have recently published structural models of the nuclear pore complex, which were produced by combining together diverse biochemical and biophysical data of different resolution levels [3].

The goal of this work is to combine experimental protein - protein interaction data of various accuracy and computational interaction modeling techniques in order to produce reliable structures of large protein complexes. The produced models, on the other hand, can be used to study interaction interfaces and the amino acids involved in binding. These predictions can be validated with experimental methods and the obtained information used to further improve the modeling in an iterative process. The main idea is to build large protein complexes from structures of individual subunits with the help of various constraints. Molecular docking was used to obtain binary protein - protein structures. Then the pairwise structures were combined using experimental and computational constraints for filtering/decreasing the combination possibilities. Many structural alternatives of the binary structures were considered which leads to a high computational complexity. High-performance computing was used to handle the computational costs. Molecular dynamics simulations were used to calculate binding energies at various stages of the assembly process.

The developed approach was applied to 3-dimensional modeling of Set1 histone methyltransferase complex from Saccharomyces cerevisiae. Set1 complex methylates lysine 4 of histone 3 and, hence, is involved in the regulation of gene expression. Experimental high-quality data set produced by protein tagging and tandem affinity purification (TAP) in combination with mass spectrometry as well as bacterial two-hybrid method formed constraints for the modeling process [4]. An ensemble of structural models were obtained for this 8-subunit protein complex.


33. Determining protein-ligand 'interaction space' in the druggable portion of Mycobacterium tuberculosis proteome.

Kalidas Yeturu and Nagasuma Chandra. Indian Institute of Science, India.

Tuberculosis remains a major health concern. Here we present a proteome level characterization of the binding site structures of mycobacterial proteins, and their comparison leading to their classification. This study provides enables binding site architecture typing and also serves as a basis for rational and systems-level approach for understanding pharmacodynamics.

Mycobacterium tuberculosis (Mtbf) has been an important human pathogen for several decades now, since it results in nearly 8 million infections and 2 million deaths every year in the world. The problem has only become more accentuated with the emergence of drug resistant varieties and the deadly synergy of tuberculosis with HIV. About 20 drugs are currently available for treatment that work by inhibiting important biochemical processes in the mycobacterial cell such as synthesis of mycolic acid, peptidoglycan and arabinogalactan or synthesis of bacterial RNA or protein synthesis. Though these drugs are extremely important, there are several problems associated with their clinical administration such as long treatment regimens and adverse effects. This is not surprising considering that the discovery process of any of these drugs has not benefited from the knowledge of the system as a whole and has also not had the advantage of structural level information of the chosen set of targets. Lack of specificity has reportedly been a major problem, as evident from the adverse effects of several anti-TB drugs.

The quest for analyzing a living organism in its entirety, has led to the emergence of several omic disciplines. Mtbf has been one of the first few organisms to be completely sequenced, which in many ways has been a turning point in mycobacterial research. The genome itself is about 4.4 million base pairs long coding for 4294 proteins. The structures of about 228 of these ORFs, have been determined by X-ray crystallography and the structural models of
2808 of the ORFs can be obtained with high confidence through standard homology modelling methods. The genome sequencing has also triggered several systems-level studies of the microbe, including our recent comprehensive target identification study targetTB (Raman et al., 2008). This study incorporates a network analysis of the protein-protein interactome, a flux balance analysis of the reactome, experimentally derived phenotype essentiality data, sequence analyses and a structural assessment of targetability, using novel algorithms developed by us recently. targetTB thus identifies a set of a first shortlist of 773 drug targets, forming the druggable portion of the Mtb proteome, from which a set of 451 proteins have been identified as the high confidence list of drug targets.

Here we seek to explore if the binding pockets in this druggable proteome can be analysed, compared and classified in a meaningful way such that patterns about ligand binding profiles can be obtained and can ultimately be used for fine-grained target prioritization, lead identification and understanding the mechanisms of drug action as well as mechanisms of drug adverse effects. We therefore obtain all putative pockets in the 767 potential drug targets identified through targetTB, amounting to 3500 pockets. Prediction of binding sites has been carried out in two phases—first by using PocketDepth (Kalidas and Chandra, 2008) (PD) a depth based method involving centrality of a grid cells and density based clustering (DBSCAN), recently developed by us. Next, in order to consider the conservation of residues at the sites, top 3 predictions from LigsiteCSC (Huang and Shroeder, 2006) (LigCSC) have been considered and only those that were consistently predicted by both methods were considered for further analysis. An all-vs-all comparison of the pockets has been performed using Pocketmatch (Yeturu and Chandra, 2008), a binding site comparison method developed in-house and pair-wise scores for similarity of binding pockets have been obtained. Pocketmatch computes 90 different sorted lists of distances, which are represented as strings and aligned using a greedy string matching algorithm that considers both the site geometry and chemical nature of the amino acid residues at the site. The similarity score of the best scoring pair between two proteins is then used as a distance measure for clustering proteins. This procedure yields 10 different clusters. Analysis of the clusters obtained and the different types of binding site architectures will be presented. Their implications for lead identification and understanding mechanisms of drug action will also be presented. This study also helps in identifying binding sites that are not identified by available crystal structure complexes. More importantly, the classification provides a framework to understand common binding site architectures among different proteins, which in turn is necessary to understand ability of these proteins to bind common ligands. This knowledge is essential to gain insights into rationalizing the observed pharmacodynamic profiles of drugs. The methodology when extended to host-pathogen comparisons, can also be used for explaining and ultimately modifying ligand design for avoiding adverse effects.

References
Huang, B. & Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation BMC Struct Biol, 2006, 6, 19-19

34. Virosphere specific protein folds.

Aare Abroi, Estonian Biocentre. Estonia.

Virosphere is definitely an integral and important part of the biosphere. In contrast to three superkingdoms, the virosphere specific protein structure space is much less studied. In the current study, virosphere specific protein domain families were identified together with corresponding 3D domains. Preliminary analysis of these data will be presented.

A lot of different studies have been carried out to analyse the distribution and abundance of protein domain families among the three kingdoms of life. The studies have involved both sequence families and structural domains at different structural classification levels. However, in most of these investigations a very important part of the biosphere – the viruses – has been ignored. (This is not unique for structural bioinformatics, it is rather common also in other fields.) It has been reported that viruses are “most abundant biological entities on the planet” (1), with the total number of virus particles exceeding the number of cells by at least an order of magnitude (2, 3). It has been proposed that viruses are even more ancient than the common ancestor of life (4, 5). Definitely, viruses as parasites have shaped the evolution of cellular organisms. So it is important to analyse also the virosphere and virosphere specific (i.e. found only in virosphere) sequence and structure space. For this purpose, virosphere specific protein domain families were identified together with corresponding 3D domains. The PfamA database was used as source for sequence domain and the SCOP database for respective structure classification. During the first step of the study,
3DSIG 2009, Stockholm - June 27-28, 2009

530 virosphere specific domain families were found, and for 87 of them the 3D structure is known. On average, the structure is known for ~24% of the PfamA domains. ~11% of virosphere specific domains are covered by structures and thus are clearly underrepresented in PDB (like in other kingdom specific domains where the structure is known for ~13–16% of domains). For example ~59% of the domains found in all kingdoms are covered with structures.

Considering two statements: a) the structures are more conserved than sequences; and b) viruses evolve much faster then their host, the following question arose - at which structural classification level these domains are virosphere specific. The structures of virosphere specific domains were analysed using the SCOP database. Using this database, the highest virospshere specific structural classification level was identified. The domains at this level and “children” contain only viral domains, and the next level “up” (or „grand”) contains also non-viral domains. According to this analysis, 39 domains were virosphere specific at “fold”, 33 at “superfamily”, 11 at “family” and 8 at “species” level. The current release of SCOP includes 1037 “folds”, 1685 “superfamilies” and 3360 “families”. So, the virosphere specific domains are enriched with a „fold” level compared to other levels. Virus specific folds were found in all major functional classes of viruses. The viral capsid/coat proteins are bona fide virosphere specific. There still remain 16 virosphere specific “folds” which are not related to coat/capsid proteins. The existence of virosphere specific domain “folds” arise the question on the origin of these structures. Several possibilities can be envisaged: a) ancestral viral folds; b) “de novo” or “ab initio” folds “generated” by viruses; c) folds leased from cellular organisms but diverged too fast and thus are not recognised by sequence analysis; d) ancestral lease from cellular organism but later lost by cells. More detailed structural bioinformatics study is required to distinguish between them.

The fold class distribution (alpha, beta, a+b, a/b, small, multidomain) of virosphere specific domains was also analysed. In virosphere specific domains, all fold classes were found. However, if we include into the analysis only the domains which are virosphere specific at “fold” and/or “superfamily” level, the a/b class is absent. It is interesting to note that in two papers it has been proposed that a/b “fold” is most likely older (or at least most abundant before the divergence of the three kingdoms) than other fold classes (6, 7). The virosphere specific “folds” exist and it would be very intriguing to understand their origin.


35. Assessing conformational diversity in proteins using evolutionary information.

Nicolas Palopoli, Ezequiel Juritz, Sebastian Fernandez Alberti and Gustavo Parisi. Universidad Nacional de Quilmes, Argentina.

We have previously presented a method for discriminating among protein models based on our SCPE program. While Now we show that it is able to recover not just a unique minimum-energy structure, but now we show that it is able to recover a complete set of native-like models representing the substitution pattern of the target protein.

SCPE algorithm

It is well established that the native state of a protein comprises a number of structures which accounts for the conformational diversity found in a protein. Most of the current methods in decoy selection after comparative or ab initio modeling procedures are not able to properly recover this group of structures that characterizes the native state.

We have presented a method for protein model discrimination based on structural features combined with evolutionary information (Palopoli et al, 2008). It is based on SCPE (Structurally Constrained Protein Evolutionary model; Parisi and Echave, 2005), a model which simulates protein evolution under structural restrictions.

Here we show that SCPE is very sensitive to backbone and sidechains 3D variability. After SCPE runs and maximum-likelihood calculations, different scores were derived to assess the 3D quality of a set of protein models obtained from recent CASP experiments. The same protocol was applied to a set of decoys built through normal modes analysis (NMA) for the same CASP protein targets. These decoys globally represent the conformational space of the native ensemble. We found that the scores for the best decoys selected using our method overlap with the scores of the decoys obtained using NMA, allowing us to select structural models belonging to the conformational ensemble of a protein native state. This method would be of great help in protein model prediction and
validation and also, it would make possible the subsequent
detection of protein functional promiscuity.

36. Computational Characterization of Human Serum Paraoxonase’s Structure and Its Interaction with VX.


Human serum paraoxonase (HuPON1) is a potential bioscavenger for organophosphorus nerve agents such as VX. HuPON1’s 3D structure and interaction mechanism with VX are unknown. This study computationally characterized HuPON1’s 3D structure, and its binding mechanism for VX. Key active site residues and associated functions were determined from the results.

Human serum paraoxonase (HuPON1) is a potential bioscavenger for organophosphorus (OP) nerve agents such as VX. A previous study suggests that increasing HuPON1’s hydrolysis activity towards VX 10-fold will make the enzyme an effective countermeasure against this nerve agent. Currently, HuPON1’s three-dimensional structure and interaction mechanism with VX are unknown. This study aims to computationally characterize both HuPON1’s 3D structure, and its binding mechanism for the VX nerve agent. Three-dimensional structures were determined through two homology modeling approaches. Bound HuPON1-VX conformations were then characterized using static binding software (DOCK, AutoDock, and OpenEye) in combination with molecular dynamics and steered molecular dynamics. Key active site residues and their associated functions were next determined from the binding results. The key residues are identified as: E53, H115, N168, F222, N224, L240, D269, I291, F292, and V346. Overall, the results provide important information on how HuPON1 initially sequesters the VX nerve agent prior to catalysis, and which residues are essential for this interaction. The binding mechanisms also provide some insight into potential catalytic mechanisms.

37. Structure-based prediction of antibody epitopes.


We present a novel method for antibody epitope prediction in protein antigens of given structure. The method uses sequence and structural properties of epitopes known from 3D structures of antibody-protein complexes and supervised machine learning. The performance of the method exceeds other available structure-based methods for antibody epitope prediction.

Reliable prediction of antibody, or B-cell, epitopes remains challenging yet highly desirable for the design of vaccines and immunodiagnostics. When a whole protein, pathogenic virus, or bacteria enters into the organism, the pathogen’s antigens are recognized by immunoglobulin receptors on the surface of B cells that respond by producing antibodies specific for these antigens. Antibody binding patches on the surface of protein antigens, known as discontinuous, structural, or conformational epitopes, are difficult to predict or identify from functional assays without knowledge of a three-dimensional (3D) structure of a protein [1]. However, limited availability of known antibody-protein structures impeded the progress in developing structure-based method for antibody epitope prediction: only a few methods are available [2-4], and their performance remains to be improved [5].

In this work, we asked whether a comprehensive analysis of sequence and structural properties of 3D structures of antibody-protein complexes enables reliable prediction of epitopes or this task is still infeasible given our current knowledge of protein antigenicity and our antibody repertoire. To carry out such an analysis, we started from extending the representative set of proteins with known structural epitopes [5]. Filtering out variable fragments of T-cell receptors and antibodies, small and very loose proteins, we selected a training set of 42 proteins for which only one epitope was
known and a test set of 17 proteins with 38 epitopes. For proteins in both sets, we generated the training and test sets of epitopes and non-epitopes. In doing so, we assumed that an epitope is represented by protein surface residues inscribed in the circle of radius R with a center at the Calpha atom of the residue that is the center of mass of an actual epitope, which was defined as residues consisting of atoms that are at 4 Angstroms distance from atoms of an antibody. At most ten non-epitopes were generated for each protein on multiple tries, taking random Calpha atoms of the surface residues as centers of circles such that non-epitopes shared at most half of the residues and excluded residues of the actual epitope. The average radius, R, at which the generated epitopes contained 90 percent of the residues of the actual epitope was equal to 15 Angstroms. For R varying between 5 and 20 Angstroms, we calculated more than thirty sequence and structural properties of epitopes and non-epitopes, using ICM [6] and ConSurf [7]. The ability of descriptors to discriminate epitopes from non-epitopes in the training set was evaluated using Kolmogorov-Smirnov, Wilcoxon signed-ranked, and t-test statistical tests. Further, the descriptors significant for epitope/non-epitope discrimination were used for training a number of supervised machine learning algorithms available in RapidMiner [8]. Finally, the best classifiers were evaluated on the test set.

Among properties discriminating epitopes from non-epitopes were propensities of polar and charged amino acids, beta-strand and coil secondary structure, and evolutionary conservation score. The best classifier was Naïve Bayes. At 5-fold cross-validation and an R of 15 Angstroms it gave an AUC (area under ROC curve), accuracy, Spearman correlation coefficient, sensitivity, and specificity values of 0.86, 86%, 0.57, 66%, and 90%, respectively. On the test set of 17 proteins the classifier showed an accuracy of 85%, correctly classifying 53% of 38 epitopes and 93% of 190 non-epitopes. No epitope was recognized in eight proteins tested. Among them were both monomers and homodimer of CIC chloride and potassium channels, which contain long alpha-helices and cannot be crystallized on their own, an insulin receptor ectodomain and extracellular region of epidermal growth factor receptor 2, which are both loose, non-globular proteins. While we expected these latter proteins to be hard to predict, the method’s inability to find epitopes in other proteins might reflect the over-fitting of non-epitopes during the learning -- non-epitopes are likely wrongly predicted because of a lack of specific knowledge of epitopes within this class of protein. However, in comparison with other structure-based methods [2-4] the performance of a proposed method is deemed superior, and further improvement appears feasible as more structures of antibody-protein complexes become available.

References

38. FREAD revisited: template-based loop prediction.

Yoonjoo Choi and Charlotte Deane. Oxford University, UK.

Over recent years, template-based loop structure prediction methods have mostly been ignored. In light of the continued rapid expansion in the number of known protein structures we have re-evaluated and improved FREAD, a database search method. In a direct comparison to MODELLER and other ab initio methods, the improved FREAD is found to perform significantly better, provided that a suitable template fragment can be found.

Loop structure prediction remains one of the bottlenecks in the 3d modelling of proteins. There have traditionally been two main approaches to loop structure prediction: database search (template-based) and ab initio. Over recent years, database search methods have mostly been ignored. In light of the continued rapid expansion in the number of known protein structures we have re-evaluated and improved FREAD, a database search method we originally developed in 2001. FREAD searches the Protein Data Bank (PDB) for fragments matching the loop region using four main filters:
1. C-alpha separation (distance in space)
2. Environment-specific substitution score (sequence-based score)
39. Assessing the Stability of Protein Complexes within Large Assemblies.

Frederic Cazals and Tom Dreyfus. INRIA Sophia, France.

This work introduces (i) tolerated collections of balls to represent protein assemblies known with uncertainties (ii) a method to highlight stable complexes within such assemblies. This method is used to check the coherence between TAP data and plausible reconstructions of the Nuclear Pore Complex.

Large assemblies and tandem affinity purification data: complexes or mixtures?

Structural genomics projects, in particular those exploiting Tandem Affinity Purification (TAP), have revealed remarkable features of full proteomes [G06]. While these insights are essentially of combinatorial nature, that is a number of proteins are known to interact within a complex, leveraging this information will require building three dimensional models of these assemblies. Such an endeavour has recently been completed for the Nuclear Pore Complex, for which plausible reconstructions have been computed from different experimental data, including TAP data [A07a,A07b]. The reconstruction procedure used is reminiscent from NMR in that restraints encoding the coherence w.r.t. experimental data have been used to defined the optimization criterion. Yet, a full synergy between TAP data and the 3D reconstruction is not at play for two reasons. First, the models built are qualitative, that is no resolution whatsoever is claimed. Second, the connexion between the reconstruction and the TAP data has not been elucidated. In particular, deciding whether proteins seen on a gel of a TAP experiment correspond to a single complex or a mixture of complexes within the assembly is not addressed. This work provides methods answering both questions.

Tracking stable complexes in uncertain assemblies.

Consider a large protein assembly. As advocated in [A08], because of the uncertainties attached to the experimental data used to reconstruct such an assembly, the resulting models are likely to be uncertain. To account for these uncertainties, we propose to use tolerated models, such a model being a collection of balls whose radii are not fixed but instead span certain ranges. More precisely, a tolerated ball is defined by a triple (c, rmin, rmax), with c the center, rmin and rmax the extreme values for its radius, with rmax>=rmin. That is, given a real number l in [0,1], we may interpolate between these values to define r(l)=rmin+lrmax-rmin. The ball of radius r(0)=rmin (r(1)=rmax) corresponds to the minimum (maximum) geometry. We define a tolerated protein by a collection of tolerated balls, and an tolerated assembly by a collection of tolerated proteins. Since such an assembly has a continuum of possible geometries, we wish to select its stable features. To do so, imagine all balls are grown simultaneously by letting the parameter l span the interval [0,1]. Typically, balls which are disjoint at l=0 merge at some point. Along the growth process, we track the evolution of the connected components (c.c.) of the union of balls. More precisely, a c.c. has a birth date lb (when two or more balls merge) and a death date ld (the c.c. merges with another one), so that a quantitative measure of the stability of the c.c. is given by ld-lb. Those which are long lived are the stable features of the assembly. An overall view of the growth process is encoded in a special graph called a Hasse diagram, which is a forest of trees, the length of a branch of a tree being precisely the quantity ld-lb of some c.c., that is, of a complex. Technically speaking, the growth process just described is associated to a so-called multiplicatively-additively weighted Voronoi diagrams--CW-diagram for short. As opposed to the Voronoi diagrams used so far in molecular modeling, which are power i.e. affine diagrams, the bissectors of a CW-diagram are in general degree four algebraic surface patches. Since we are not aware of any algorithm able to construct such a diagram, we track the variation of the c.c. by probing this Voronoi diagram with a power diagram. The probing process consists of hunting the value of parameter l yielding a change in the number of c.c.

Application to uncertain assemblies reconstructed from tAP data.

The construction just sketched can be used to disambiguate TAP data. Assume that we are given a tolerated assembly, consisting say of n tolerated proteins (instances in the sequel) of p different types (types in the sequel). For example, for the NPC, one has n=456 and p=30. Consider now a set Q of q types (q<=p) seen in a TAP experiment. By running the previous machinery on balls of the prescribed types in the tolerated assembly, we can state whether the set Q corresponds to an isolated complex, or to a mixture of (stable) complexes.

A case study: the nuclear pore complex.

In [A07a,A07b], 1000 optimized structures of the NPC have been reconstructed from various experimental data including ultra-centrifugation data, cryo-EM, immuno-EM, overlay assays, and TAP data. These plausible structures have also been merged to define an average structure, which is of qualitative interest. Indeed, quoting the authors: “Our map is sufficient to determine the relative
positions within NPC; we do not interpret features smaller than this precision" [A07a]. Starting from this mean structure, we show how to build a toleranced model of the NPC. Running our characterization of stable structures on this model provides a quantitative answer for deciding whether TAP data correspond to single complex or a mixture of complexes. This work has been carried out for the 83 pullouts of [A07a]. The example of composite #14, which contains Nup84 and Nup145C, is provided on Fig A. The left panel shows a top view of a mixture of 16 complexes (8 on each side on the symmetry plane of the NPC), while the right panel provided the aforementioned Hasse diagram.

CONCLUSION

This work presents a method allowing one to make quantitative assessments for uncertain / tolerated models. In particular, it allows a structural interpretation of TAP data in conjunction with structural models of low to intermediate resolution. We believe this is an important step to build mechanistic models of large assemblies, so as to start investigating their dynamics.

REFERENCES


Aquaporins allow the diffusion of water and small solutes across biological membranes. We calculated the channel electrostatics for nine 3D-structures representative of different aquaporin subfamilies, finding that specific electrostatic profiles correspond to the main selectivity to water or glycerol and that individual pore-lining residues strongly affect the channel electrostatics.

Aquaporins are homotetrameric channel proteins, which allow the diffusion of water and small solutes across biological membranes (Preston et al, 1992). Aquaporins have been discovered in all the domains of life, from bacteria to mammals. They are characterized by two highly conserved Asn-Pro-Ala (NPA) sequence motifs, located at the end of two half helices which lie at the middle of the permeation channel, where they form a constriction. Another narrower constriction, known as aromatic/Arg (ar/R) selectivity filter, is found at the periplasmic side of the channel and is formed by four residues including aromatic residues and a widely conserved Arg.

In humans, aquaporins are involved in several diverse functions, like concentrating urine in kidneys and participating in forming biological fluids, and a number of diseases, like nephrogenic diabetes insipidus, congenital cataracts, Meniere’s disease, cerebral edema and stroke. Moreover, protozoan aquaporins, which are common components of the parasite-host interface, play essential roles for the life of the parasitic organisms and have been addressed as potential targets for chemotherapy in the treatment of malaria (Beitz, 2005).

Aquaporins have been grouped by phylogenetic analyses into about thirty major subfamilies, characterized by their specific transport selectivity and efficiency, source, localization and physiological function (Zardoya, 2005). According to their main transport selectivity, aquaporin subfamilies can be further classified into “orthodox aquaporins”, which allow the flux of water molecules only, and “aquaglyceroporins”, which facilitate the diffusion of glycerol and other small solutes in addition to water. Mechanisms of selectivity in orthodox aquaporins and aquaglyceroporins and role of individual residues in the pore are not yet fully understood. Currently, experimental 3D-structure representatives of 9 different aquaporin subfamilies are available, that include water and glycerol channels and the bi-functional protozoan PfAQP (Newby ZE, 2008). This variety of aquaporin structures represents a valuable resource for understanding the structure-function relationships within this protein family. To gain insights into aquaporin selectivity to either water or glycerol, the continuum Poisson-Boltzmann electrostatic potential along the channel was calculated and compared for the nine 3D-structures which are representatives of different aquaporin subfamilies, and a panel of functionally characterized mutants for which high-accuracy 3D-models could be derived. Interestingly, specific electrostatic profiles associated with the main selectivity to water or glycerol could be identified. Moreover, evaluation of electrostatics of the mutants, along with a thorough sequence analysis of the aquaporin pore-lining residues, illuminated the contribution to the electrostatics of individual pore-lining and “second-shell” residues. Therefore, we propose that electrostatic profiles of aquaporin channels are potentially highly predictive for the selectivity of new uncharacterized aquaporins, and may be used to assist the design of aquaporin mutants, prior to functional characterization.

Beitz E. (2005), Aquaporins from pathogenic protozoan parasites: structure, function and potential for chemotherapy. Biol Cell, 97, 373-383
Zardoya R (2005) Phylogeny and evolution of the major intrinsic
protein family, Biol Cell, 97, 397-414

41. Molecular evolution of class A protein-coupled receptors probed by the proline pattern of transmembrane helix 2.

Julie Devillé, Julien Rey and Marie Chabbert. UMR CNRS 6214 - INSERM U771, France.

Class A G-protein-coupled receptors possess an unusual proline pattern in helix 2. Sequence analysis, 3D data mining and molecular modeling indicate that it is related to an indel arisen in a bilaterian ancestor. This proline pattern is an efficient marker to probe molecular evolution of GPCRs.

Class A G-protein-coupled receptors (GPCRs) constitute the largest transmembrane receptor family of the human genome. These receptors possess about 4% of highly conserved residues, forming characteristic motifs. Helices 5, 6 and 7 possess a highly conserved proline residue. However, helix 2 presents an intriguing feature with a variable proline pattern. Proline may be located at position 2.58, 2.59, 2.60 or may be absent from helix 2. The ratio of these different patterns depends upon the species considered. In particular, the weight of P2.58 receptors increases from 7% in C. elegans to 40% in H. sapiens, suggesting that the proline pattern of helix 2 might be used as an evolutionary probe.

3D data mining and molecular modeling were carried out to investigate the origin of this unusual proline pattern [1]. In the crystallographic structures of class A GPCRs presently available [2-5], helix 2 is characterized by a pi bulge related either to a G2.56G2.57 motif (rhodopsin) or to a P2.59 pattern (beta2AR, beta1AR, A2AR). To determine whether this bulged structure is consistent with different proline positioning, we searched for structures similar to the known structures of helix 2 and analyzed the position of proline in the hits. The search was carried out with the SPASM program in a non-redundant subset of the Protein Data Bank (4900 protein chains) [6]. In about half the hits, a proline residue was present at position equivalent to position 2.59 or 2.60 of the query. The search motifs based on beta2AR, beta1AR and A2AR favored hits with proline at position equivalent to 2.59, whereas those based on rhodopsin favored hits with proline at position 2.60. This difference arises from the local structure of the bulge in the search motif. No hit possessed a proline residue at position equivalent to 2.58, indicating that proline at this position is highly disfavored in bulged structures similar to helix 2.

The next step was the modeling of a P2.58 receptor with deletion of one residue in the bulge elbow of helix 2. The chemokine receptor CCR5 was chosen as an example of P2.58 GPCRs as it is widely studied in the literature for its role as HIV co-receptor. Using either the rhodopsin or the beta2AR structures as templates, only minor reorganization of the seven-helix bundle (rmsd lower than 1.3 angstroms) was required to accommodate the deletion in helix 2. The Protein Data Bank was searched for structural motifs similar to the modeled helix 2 of CCR5. The high propensity of proline at position 2.58 in the hits (7.5) indicates that the modeled helix 2 is similar to proline-induced kinks commonly found in protein structures.

These data strongly support the assumption that a similar general organization of helix 2 can be reached either through the bulged structure of helix 2 that is observed in rhodopsin and beta2AR or through a “typical” proline kink, resulting from deletion of one residue in the bulge. The bulged structure of helix 2 should be present in receptors without proline or with proline at position 2.59 or 2.60, whereas the kinked structures should correspond to receptors with proline at position 2.58. These two structures are related by an indel. This indel induces mainly local structural changes and should not affect the overall fold of the seven helix bundle.

We investigated the history of the proline pattern of helix 2 during evolution, by analyzing the sequences of class A GPCRs from five species: C. elegans, D. melanogaster, C. intestinalis, D. rerio and H. sapiens. In the most ancient species, C. elegans, P2.58 receptors can be assigned as somatostatin/opioid receptors. These receptors are highly related to the galanin/kisspeptin receptors that possess a P2.59 pattern and belong to the peptide receptor sub-family. This strongly suggests that the indel in helix 2 arose very early in GPCR evolution, in a bilaterian ancestor, and led to the split between the P2.58 somatostatin/opioid receptors and the other P2.59 peptide receptors. Sub-families with proline at position 2.59 or no proline expanded earlier, whereas P2.60 receptors remained marginal throughout evolution. P2.58 receptors were evolutionary successful in vertebrates with the rapid expansion of the chemokine and purinergic receptor sub-families from somatostatin/opioid related ancestors. These receptors are involved in the regulation of the vascular and immune systems specific of vertebrates. The proline pattern of helix 2 provides an efficient marker to probe molecular evolution of class A GPCRs [1].

42. Prediction of causative effects of disease-related mutations at the molecular level by integrated bioinformatic analyses.

Jan Kosinski and Janusz Bujnicki. Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Poland.

We present an integrated approach to predict the effects of missense mutations by mapping the functional sites using a broad spectrum of bioinformatic methods, such as sequence and structural analyses, modeling and docking. Our approach can be easily applied by biologists and serve as a guide for selecting appropriate tools.

Explaining the effects of disease-related mutations at the molecular level could allow rational designing of therapies specific toward particular molecular dysfunctions. Here, we apply an integrated approach to predict the effects of missense mutations by mapping the functional sites using a broad spectrum of available bioinformatic methods.

Our procedure starts with a detailed sequence analysis and ends with structural modeling and comprehensive structure analysis. We apply several methodologies such as mapping of sequence conservation, prediction of protein-protein and protein-DNA interaction sites, prediction of small ligand binding sites, conformational analysis using normal mode analysis, fold recognition, comparative and de novo structural modeling, and computational protein-protein and protein-DNA docking. Using state-of-the-art tools and alternative programs and servers, we then derive consensus prediction of functional regions in the protein. This, in turn, allows for classifying the known disease-related mutations based on the molecular function with which they could interfere.

We show the effectiveness of our approach on the example of the C-terminal domain of human MLH1 protein (MLH1-CTD). About 30 missense mutations found in MLH1-CTD have been associated with HNPCC syndrome, which increases the risk of development of several types of cancers. Our integrated analyses allowed predicting a dimerization site, two protein-protein interaction sites, a DNA binding site and a hinge region most likely responsible for functionally important movement of two subdomains of MLH1-CTD. The predicted sites allow classifying the HNPCC-related mutations in MLH1-CTD based on the molecular function with which they presumably interfere. Some of our predictions were already confirmed by experimental analyses.

As our integrated approach uses only publicly available software (mainly web servers) it can be easily applied by biologists. Our work presented here could then serve as a guide for selecting appropriate methods and particular tools.

43. STP: A Program to Predict Protein Binding Surfaces.


We present a novel algorithm, STP (Surface Triangle Profile) to predict protein-ligand binding sites. This algorithm uses the chemical composition of the surface of the protein to assess the likelihood of a certain patch at the surface to be a binding site.

Biological function is a result of intermolecular interaction. Proteins interact with other proteins, peptides, sugars, or other small molecules to create functional complexes. These interactions are vital to molecular functions and have proved to be the backbone of drug design theory. Locating binding sites on protein surfaces is a very important step in studying such interactions. By locating the binding site of a certain protein, it becomes possible to conduct virtual screening experiments to find a suitable inhibitor/regulator. Locating the binding site also gives way to introducing targeted mutations aimed at regulating/blocking a certain interaction.

The method presented in this work is based on creating a score table that would indicate the propensity of surface atoms to exist in ligand binding sites. By focusing on atoms, the noise created by profiling residues is avoided. STP classifies protein atoms according to the 13 Atomic Groups[1] to predict protein-ligand binding surfaces. In order to search for specific patterns on the surface, water accessible atoms are grouped in triplets (a triplet being a...
group of 3 surface atoms that can be simultaneously touched by a water probe), giving rise to a total of 455 distinct triplets, also referred to as triangles (Figure 1A).

Patches on the surface are scored via a sliding window scheme. Each patch is given the average score of all triangles it contains, and recorded at its center atom. Those scores are then normalized to a scale of 0-100 and stored in the B-Factor column of the PDB file (Figure 1B). This enables most molecular viewers to color the protein structure on a temperature color scale (blue to red).

The power of STP has been evaluated by comparing it to several algorithms. While it is common knowledge that the deepest cleft on the surface of a protein is probably the binding site of that protein, we have showed that the cleft with the highest STP score is a better estimate of the location of the binding site. Furthermore, the performance of STP was compared to that of Q-SiteFinder and the method created by Morita et al. (2008)[2]. The test was run on a dataset of 35 structurally distinct proteins in the unbound state which share structural similarity with 35 proteins in the ligand-bound form (which was created by Laurie and Jackson[3]). STP was run to locate the binding sites on the unbound proteins. The unbound proteins were then superimposed onto their bound homologues. Ligands were then extracted to mark the binding sites in the unbound proteins. The cavities on these proteins were calculated with SURFNET[4] and then ranked by their STP score. Although those 2 methods had the advantage of having the bound state of some of the test structures in their training set, STP still outperformed both methods.

This work outlines a novel method of studying protein interaction. Our results indicate the importance of the chemical characterization and annotation of surface atoms. The influence of these surface atomic patterns on predicting protein function is yet to be studied.


44. Organic enzyme cofactors – from structure to function.

Julia Fischer, Gemma Holliday and Janet Thornton, European Bioinformatics Institute (EMBL-EBI), UK.

Many enzymes not only require the 20 amino acids to catalyse the chemistry of life, but also cofactors, metal ions as well as organic cofactors. Here we present a study of the 3D structure conformations of organic enzyme cofactors and their relation to cofactor function.

Enzymes catalyse all the metabolic reactions the cell requires for survival. The only tools required to enable the full chemistry of life are the 20 amino acids, some inorganic cofactors like metal ions and organic enzyme cofactors. While the chemical function of amino acids [1] and metal ions [2] in enzyme mechanisms continues to be studied, the role of organic enzyme cofactors in biocatalysis has yet to be studied in depth. In this work we will examine how organic cofactors assist in enzyme catalysis. At the meeting, we will present our latest results on the structure-function relationship of organic enzyme cofactors as well as the results of our mode-of-binding analysis. We will discuss the relevance of this approach for biological research and implications for drug design.

Three-dimensional structures of cofactors are often crystallised alongside enzymes and therefore are publicly available from the PDB. We use these data to investigate the conformational variability of enzyme cofactors when bound to a protein and relate the structural clusters to functional states of the cofactors. As methods, we use structural superposition of cofactors to assess the conformation space for each of the molecules. We perform a structural clustering [3] to obtain structurally similar groups. The cofactor clusters are then associated with the E.C. numbers, CATH domains, and species of the original protein 3D structure. E.C. numbers hierarchically describe the chemistry of the overall enzyme reaction and thus may allow us to detect relations between cofactor conformation and chemical function. However, cofactor function may remain the same in different E.C. numbers. CATH domains hierarchically classify the 3D structure of the protein domain and allow us to link cofactor conformation with domain evolution. Finally we aim to refine our evolutionary analysis by investigating possible links between cofactor conformation and the species in which they occur.

In the second part of the work, we analyse the mode of binding of the cofactors in proteins. This approach enables us to assess the way that cofactors are bound to domains and how this relates to chemical function and evolution. We assess the mode of binding for each atom of the cofactor using relative solvent accessibility (RSA), which measures the percentage of the surface of an atom or molecule that is accessible to solvent. To realise this analysis, we
employ the NACCESS [4] method to compute the RSA of the whole cofactor in the biological assembly, as well as per atom. We then compute the average and the variance in RSA over all instances of the same cofactor. The overall RSA of the molecule, averaged over all structures of one cofactor, shows how buried the ligand is in the protein. The RSA per atom provides insights into which part of the cofactor is always buried and which is more accessible. A low variance in RSA indicates that this feature is important for function because it is conserved over many structures.

We will present interesting examples from these analyses and discuss biological implications. One of these examples is Vitamin B12: the conformations of the 13 structures are clustered into four groups. In group one (Fig. 1, shown in yellow), the cofactor is in the base-on state, in group 2 and 3 (light pink and dark pink) it is in base-off/His-on state and in group 4 (green) it is in base-off state. These three states are characteristic for the catalytic state of the cofactor [5]. Normally, a nitrogen atom of the endogenous ligand, 5,6-dimethylbenzimidazole is a ligand of the cobalt atom, which is located in the centre of the porphyrin ring (base-on state). When bound inside the target enzyme, the base is often replaced by a histidine residue, which is provided by the target enzyme. Lately it has been found that in humans, the cofactor also exists in a base-off state, which means that the 5th coordination position is empty. This conformation of B12 is found in adenosyltransferase, the enzyme that transfers the adenosyl residue onto the cofactor and delivers it to the target enzyme methylmalonyl coenzyme A mutase [5]. In this case, the structural conformation of the cofactor therefore directly corresponds to one of three possible functional states: base-on (catalytically inactive), base-off/His-on (catalytically active) and base-off (not yet in target enzyme).

Fig. 1: Structural superposition of vitamin B12: base-on state (yellow), base-off/His-on states (light and dark pink) and base-off state (green).

References:

We present an animation system based on the 3D open-source software Blender applied to the transition between different conformations of Calmodulin. Results are compatible with experimental data, and are visualized using a technique that reveals chemical and physical features immediately. The instrument can be used in research and other settings.

The fact that proteins move continuously in the cellular environment has finally overcome the concept of structure-function, which tacitly implied a single conformation for any given molecule. The features of the movements, however, are still difficult to calculate and represent, in particular for practicing laboratory (wet) biologists, who are the people that would most benefit form an easy visualization of such motions. Beside the change in conformation, resulting in different shapes of the molecules, relative motion of the aa composing the surface of the proteins confer dynamic properties to the surface, such as the exposition of hydrophobic patches, or availability of charged spots.

In an effort to address both these aspects (motion and visualization) of moving proteins, we have implemented a series of python scripts that work in Blender, an open source program of 3D animation, visual effects and video games.

A system was devised that allows a (relatively) fast calculation of the conformational changes. A script to import and export multiple models from pdb data in the 3D scene has been written, and molecular transitions are achieved by using internal Blender game engine, equipped with a set of rules that mimic chemical behavior. Intermediate conformations so generated are automatically adjusted using the force field of Swiss-PDBviewer and compared (RMSD) with other conformations derived from the original NMR collection. With a few cycles of this procedure we have generated a continuous motion that covers a geometrical distance of 21 Å in steps of <1 Å, ‘visiting’ a dozen conformations experimentally derived.

The movement was visually inspected and it shows a smooth motion which positively interacts with the aesthetic sense of all viewers tested, thus responding to our innate sense of harmony typical of equilibrated biological entities (i.e. it’s really beautiful). The biological and biochemical meaning of this motion can now be investigated.

45. Protein motion and visualization using game engine.

For visualization of chemical and physical properties, we have developed a visual code applied to the surface of the molecule without using colors.

Each frame from the animation is saved and exported as a pseudo-pdb file, converted to .pqr and elaborated with and a series of libraries and other programs (VMD) and scripts (pyMLP.py) that calculate electrostatic and lipophilic potential values. Using a homemade program, these fields are mapped on the protein surface (obtained as an .obj file from Chimera) and associated to all vertices of the mesh. These values are fed to the texture nodes of Blender and are finally expressed as scale of smooth-shiny to bumpy-dull for MLP and with particles emitted from and attracted to the surface for EP. One of the resulting images is shown in the figure: here EP particles are rendered in color because a single image cannot provide information about the motion.

These visual properties are immediately decoded by humans, as they correspond to textures of real material (MLP) and to familiar dynamic behavior. We show results obtained using NMR files 1cfc and 1x02, which contain 25 and 20 conformations of Calmodulin, Calcium-free and Calcium-loaded respectively.

47. Bioinformatics for GPCRs oligomerization


The formation of GPCRs homo and hetero oligomer has been suggested by biochemical and pharmacological evidence. To clarify the mechanism of signal transduction, we have developed a method to predict interfaces for GPCRs oligomerization (GRIP : http://grip.cbrc.jp/GRIP/index.html). We will introduce our recent works associated with GPCRs oligomerization.

Introduction

G-Protein Coupled Receptors (GPCRs) are one of the most important pharmaceutical targets. They function not only as monomers but also as dimers or higher-order molecular complexes. Recently, GPCR oligomerization and its functional meanings have been extensively investigated [1][2]. These studies have revealed a wide variety of biochemical functions of the oligomers in cells, the combinations of subtypes required for complex formation, and the biological functions of the oligomers [1][2]. Nevertheless, many questions about GPCR oligomerization still remain [1][2]. One of the strategies to answer such questions is the regulation of GPCR oligomerization by identifying and modifying the interfaces for GPCR-GPCR interaction, although few experimental studies have identified the interfaces. Therefore, it is important to predict the interfaces for protein-protein interactions between GPCRs. However, the prediction of the interfaces of GPCR oligomerization based on residue conservation is difficult, because the various subtypes of GPCRs often use different regions of the three-dimensional structure as the interfaces, even when the subtypes belong to the same subfamily. Therefore, we developed a novel method to predict the interfaces for GPCR oligomerization, based on the spatial distribution of conserved residues for each subtype [3]. We also constructed a web service to predict the interfaces for GPCR oligomerization. This service, named G-protein coupled Receptors Interaction Partners (GRIP : http://grip.cbrc.jp/GRIP/index.html), is available through the internet (Figure 1). As far as we know, it is the only web service that predicts the interfaces for GPCR oligomerization.

Method

The method is made of six steps. At first, a multiple sequence alignment of the GPCR sequences of interest is aligned together with the sequence of bovine rhodopsin whose three dimensional structure is available in atomic resolution. Second, conservation score is calculated at each alignment site. Third, calculated conservation scores are assigned to the corresponding sites of the bovine rhodopsin structure. Fourth, inner residues, extracellular loops, and intracellular loops of bovine rhodopsin are ignored because the lipid facing side of membrane spanning helices are considered to be involved in the oligomer formation. Fifth, the remaining residues are projected on the plane, which is assumed to be parallel to the membrane plane. The plane is defined by the principal component analysis of all alpha carbon coordinates of the tertiary structure of bovine rhodopsin. As a result, ring like distribution of the projected residues is generated on the plane. Finally, we adopt a sector within the ring like distribution as an interface where the number of conserved residues is significantly large and an ad hoc objective function shows the maximum score. Further details are described in reference [3].

Results and Discussions

Residues on helices IV and V play the most important interaction between monomers of the rhodopsin dimer [4]. Among them, it is currently believed that 175W is a critical role in oligomerization of bovine rhodopsin. Conserved residues in the regions I, II, III, IV, V and VI (152H, 155M, 162V, 175W and 201E) belong to helices IV and V and they are included in rhodopsin oligomeric interfaces. On the contrary, conserved residues in the regions VII and VIII are included in the interface between a pair of monomer as well as between monomers. Some atoms on the interdimeric surfaces of one molecule are located within 6 angstrom from the opposite side. The observation suggests that the conserved residues are clustered at the interface between bovine rhodopsins and that our method can efficiently detect the clustering. In the same way, predicted regions...
by our method corresponded to the experimentally suggested interfaces for other GPCRs oligomerizations [3]. Thus, the detection of the spatial cluster of the conserved residues would be useful to predict the interface. Since Class A type GPCRs have high sequence similarity with rhodopsin, precise model structures could be made. As described in some articles, however, the oligomerization patterns are different among the subtypes, despite the high sequence and structural similarity. By applying our method, the interfaces of Class A GPCRs could be predicted.

Future work
We are now developing a database for GPCRs oligomerization. The database stores information about (1) interaction partner proteins for GPCR oligomerization, (2) experimentally suggested interfaces for the oligomerization and (3) predicted interfaces on GPCR monomers predicted by GRIP. The database will also contribute to the research about GPCRs oligomerization.

References


Petras Kundrotas, Zhengwei Zhu and Ilya Vakser. The University of Kansas, USA.

Structural templates for 119,355 protein-protein interactions in 771 organisms currently represented in publicly available databases are analyzed. The resulting structural models and analysis tools are organized in searchable Genome-Wide Docking Database (GWIDD) at gwidd.bioinformatics.ku.edu. The results provide guidelines for 3D modeling of PPI networks.

Protein-protein interactions (PPI) play the key role in all aspects of cell functioning, from cell metabolism, to cell communication and cell cycle. The structural aspects of PPI, provided by large scale, genome-wide studies, are essential for the description of life processes at the molecular level. Although the gold standard for experimentally determined protein structures is X-ray crystallography, crystal structures of protein complexes are more difficult to obtain than those of individual proteins. Many protein interactions are transient, which makes studying them crystallographically even more problematic. Furthermore, the X-ray crystallography and other experimental techniques can provide structures only for a fraction of proteins, with the rest to be modeled by high-throughput computational techniques based on experimentally determined templates. Thus, most structures of protein complexes need to be determined by computational modeling. The variety of modeling methods and the rapid development of computational hardware enables structural modeling of PPI (protein docking) on a large, genome-wide scale. The model of two interacting proteins can be built either based on a homologous complex with known structure (homology docking) or by processing independently generated models of the monomers using various template-free or template-based techniques. In the latter case, the accuracy of individual protein structures, especially their binding sites, is essential for the success of modeling their complexes. However for the high-throughput modeling, computational methodology needs to be computationally inexpensive.

Homology docking was applied to all interactions in DIP (http://dip.doe-mbi.ucla.edu) and BIND (http://www.bind.ca) databases. Homology models of monomers were analyzed for docking suitability. The template complexes were identified by PSI-BLAST local alignment and profile-to-profile alignment of entire sequences. The best alignment was chosen from the resulting pool of alignments based on the alignment identity, coverage and the template quality criteria. The structural models of all PPI with available templates were generated by NEST. The assessment of the models quality was performed by comparing structural characteristics of the modeled interfaces with those of the crystallographically-determined structures. The distribution of structural templates and the resulting models from 771 organisms in DIP and BIND databases are presented and the observed irregularities are discussed. Previous benchmarking showed that the homology docking yields accurate models for ~15 – 20 % of PPI. The current study demonstrates that this percentage holds on average in the general case as well, however, it varies significantly in different species. This is illustrated by the representative Figure, where numbers of all interactions in DIP and BIND databases for five broadly defined types of living organisms are compared to the corresponding numbers of GWIDD entries (currently, experimental X-ray structures and homology docking models based on such structures). For convenience of the analysis, eukaryotes are split into lower eukaryotes (primitive organisms and fungi) and higher eukaryotes (plants and animals). PPI with partners from different organisms are excluded (except for the viruses where PPI are usually between a virus protein and a protein from the host organism). It is known that certain organisms are overrepresented in the existing PPI databases. For example, out of 28,664 interactions in fungi organisms, 28,301 are from baker’s yeast (Saccharomyces cerevisiae). For the overrepresented organisms, the fraction of PPI...
with available templates is usually significantly lower compared to the organisms with less known PPI. The overrepresentation occurs for model organism widely used in mutagenesis studies. To avoid such a bias, the analysis was performed for non-redundant PPI sets where PPI from the same organism with small sequences differences are excluded from the consideration. These non-redundant sets were further used to analyze conservation of PPI based on evolutionary distances between target and template sequences.

For systematic evaluation of potential accuracy in high-throughput modeling of binding sites, a statistical analysis of target-template sequence alignments produced by local PSI-BLAST alignment was performed for a representative set of 329 two-chain non-oblige protein complexes with known structure from the DOCKGROUND database (dockground.bioinformatics.ku.edu). For the majority (97%) of the target sequences there is at least one alignment containing all residues belonging to the interface of the target complex. For the high-identity alignments, the interface alignment was better than that of the highly conserved protein core. The full interface alignments were obtained even in the case of poor alignments when a relatively small part of the target sequence (as low as 40%) aligned to the template sequence, with a low overall alignment identity (< 30%). Although such poor overall alignments might be considered inadequate for modeling of the entire proteins, the alignment of the interfaces was high enough for docking. In the set of homology models built on these alignments, one third of those ranked #1 by a simple alignment identity criteria had RMSD < 5 Å, the accuracy suitable for low-resolution template free docking. Such models corresponded to multi-domain target proteins, whereas for single-domain proteins the best models had 5 Å < RMSD < 10 Å, the accuracy suitable for less sensitive structure-alignment methods. Overall, about 50% of complexes with the interfaces modeled by high-throughput techniques had accuracy suitable for meaningful docking experiments. This percentage will grow with the increasing availability of co-crystallized protein-protein complexes. The results offer guidelines for further studies in this area of structural modeling of protein-protein interactions by providing estimates for homology modeling of interfaces on the monomers.

The results of this work are implemented in the searchable Genome-Wide Docking Database (GWIDD) at gwidd.bioinformatics.ku.edu, where user can search and download models either for specific protein complexes with known structure from the Protein Data Bank for structures of proteins with Cys in alternate thiol redox states.

Protein Data Bank for structures of proteins with Cys in alternate thiol redox states.

Disulfides are conventionally viewed as structurally stabilizing elements in proteins but emerging evidence suggests two disulfide subproteomes exist. One group mediates the well known role of structural stabilization. A second redox-active group are best known for their catalytic functions but are increasingly being recognized for their roles in regulation of protein function. Structural and redox-active disulfides can be distinguished by their redox potentials. Disulfide redox potentials measured in thiol-disulfide oxidoreductases range from -120mV to -270mV [1-4]. For disulfides serving structural purposes, the redox potential can be as low as -470mV [5]. Computational methods for distinguishing the two groups would be a boon. Redox-active disulfides are, by their very nature, more susceptible to reduction than structural disulfides; and conversely, the Cys pairs that form them are more susceptible to oxidation.

In this study, we searched for potentially redox-active Cys Pairs by scanning the Protein Data Bank (PDB) for structures of proteins in alternate redox states. The PDB contains over 1,134 unique redox pairs of proteins, many of which exhibit conformational differences between alternate redox states. Our study is the first to systematically study these conformational changes in the PDB. Several classes of structural changes were observed, proteins that exhibit disulfide oxidation following expulsion of metals such as zinc; order/disorder transitions; changes in quaternary structure and major reorganisation of the polypeptide backbone in association with disulfide redox-activity. This latter group, also known as “morphing” proteins, undergo plastic deformations involving large scale rearrangements of the polypeptide backbone. Our study shows these proteins, which challenge Anfinsen’s thesis of a one-to-one mapping of sequence to structure, can be influenced by redox conditions.

Protein redox regulation is of growing interest because of its relevance to neurodegenerative diseases, cancer, diabetes and heart disease. Based on evidence gathered supporting disulfide redox activity, we propose disulfides present in alternate redox states are likely to have physiologically-relevant redox activity.

References

49. Conformational changes in redox pairs of protein structures.

Merridee A Wouters, Richard A George, Naomi L Haworth, Lina L Feng, Jason Y Liu and Samuel W Fan, Structural & Computational Biology Program, Victor Chang Cardiac Research Institute, Australia.

Redox-active disulfides are, by their very nature, more susceptible to reduction than structural disulfides; and conversely, the Cys pairs that form them are more susceptible to oxidation. In this study, we mined the

50. Using a Differential Geometry Approach for Characterizing Biological Relevant Interfaces.

Shula Shazman, Gershon Elber and Yael Mandel-Gutfreund, Technion- Israel Institute of Technology, Israel.

We present a novel method based on differential geometry for characterizing protein binding interfaces. The method takes advantages of the Gaussian and Mean curvatures of surface points. Using a Support Vector Machine with the differential geometric features, we successfully classified double-stranded DNA from single-stranded RNA interfaces with relative high accuracy.

Inferring protein function from structure is a challenging task. Nowadays, as a result of structural genomic initiatives, a large number of three dimensional structures of proteins for which biological function has not been characterized are available. Protein interactions, namely Protein-DNA, protein-RNA, protein-protein and protein-small molecular interactions play critical roles in many biological functions including gene transcription and translation, signal transduction, enzyme regulation and immune response. Since protein interactions are central to various biological processes, studying the underlying principle of these interactions is critical for the understanding of living cells. Proteins interact with their various partners via distinct regions on their surface. Though, evolution can lead sequences and structures to diverge, surface determinants that are essential for the protein function are likely to be maintained. These surfaces are expected to be characterized by unique chemical, physical and geometrical properties. Previous studies have focused on the characterizing the unique features of protein binding interface. However, distinguishing between DNA and RNA binding interface is still an enigma.

In this study we present a novel differential geometric method for the characterizing protein binding interfaces, specifically DNA and RNA interfaces. The method is based primarily on geometric surface properties, specifically the Gaussian and Mean curvatures using the IRIT modeling package1. In this approach each surface point is classified to one of eight different surface’s local geometry shapes based on the directions of the Gaussian and the Mean curvatures. The different local geometric shapes we analyzed are: Peak, Pit, Ridge, Valley, Flat, Minimal surface, Saddle ridge and Saddle valley. In order to infer the geometric shape composition of the interfaces we analyzed the distribution of the grid point related to local surface shapes. Our dataset contains two groups of binding interfaces from complexes of protein-RNA (single-strand), protein-DNA (double-strand). Using an unsupervised clustering approach we could clearly classify the different interfaces according to the relative distribution of the surface point associated with different geometric shapes. Specifically, we observed that the DNA-binding interfaces are characterized by a high density of interfaces points associated with the “valley” shape while in the RNA-binding interface are characterized by “ridges” (See Figure). Consequently, we applied a Support Vector Machine (SVM) to automatically classify double stranded DNA (dsDNA) from single stranded RNA (ssRNA) binding interfaces based on the geometric properties of the interfaces. Overall the accuracy of our classification for distinguishing RNA from DNA binding interfaces was 0.83. Furthermore we checked the complementarity of the RNA and protein interfaces and found that the protein and RNA interfaces show a mirror image, specifically for the Peak, Pit, Ridge, Valley, and the Saddle Ridge shapes. We suggest that the geometric properties can be combined with other protein properties (such as electrostatic potential) for predicting specific protein-RNA interactions.

In conclusion, we present a new approach to characterize and classify different binding interfaces on proteins, concentrating specifically on dsDNA and ssRNA binding interfaces. The method employed in this study is based on differential geometry and can describe variable geometric shapes on the interface (e.g. ridge). Our method differs from classical methods for classifying proteins surfaces as it does not consider the overall concavity of the protein surface but its specific shape. Thus, the advantage of the method is that is does not depend on the specific dimension of the ligand but on its overall geometry. We believe that combination of the differential geometry parameters with electrostatic features, used in our previous studies for nucleic acid binding classification2, will allow to uniquely identifying biological relevance molecular interfaces, specially distinguishing RNA from DNA binding interfaces.

Figure: Results of an unsupervised clustering approach for separating between dsDNA and ssRNA binding interfaces in proteins. A. Distribution of interface points between eight different local shapes of surface. Each row represents a protein binding interface and each column represents a different local shape ("other" column is for points which are not connected with a specific shape). The colors represent the percent of points which fall in each shape category, ranging from yellow (0% of the points) to red (25% of the points). Color bar is shown above. B. An example of DNA binding protein, the structure of a restriction endonuclease bound to DNA. The DNA binding interface (colored yellow) is classified as a “Valley”. C. An example of RNA binding protein, the E.coli cysteinyl-tRNA and T. aquaticus elongation factor EF-TU. The RNA binding interface (colored
yellow) is characterized as a “Ridge”.

References

51. Protein sequence and structure optimisation in one probabilistic framework.

Gundolf Schenk and Andrew Torda. University of Hamburg, Germany.

Protein structure prediction and sequence optimisation are two challenges which routinely appear to scare PhD students. We have used a fragment-based approach which turns both into a classic self-consistent field problem within a framework of descriptive statistics.

We are interested in self-consistent mean field (SCMF) methods and how they can be used for protein structure prediction and sequence optimisation. This also means formulating and building new force fields. Our method has a probabilistic model of protein sequence-structure correlation and approaches self consistency within this framework.

There is a myriad of approaches to ab initio structure prediction, usually based on a mixture of chemistry and physics at high and low resolution. Our approach is definitely coarse-grain, but with little preconception as to what is important. We have our own statistical assumptions, but then a very flexible method for finding sequence-structure relationships.

Our scoring scheme has its roots in previous work[1], where the application area has been the comparison of proteins and recognition of remote similarities. The method is based on a Bayesian classification of overlapping protein backbone fragments. From the protein database every possible fragment was extracted from each protein in a sliding window fashion. These fragments are described by a fixed number (5-7) of amino acid labels (sequence terms) and a fixed number (5-7) of dihedral angle pairs (structure terms). A maximally parsimonious automatic classifier[2] was used to cluster the fragments into classes, which model the sequence terms by multiway Bernoulli distributions and the structure terms by bivariate Gaussians. The classification is then defined by the number of classes (typically 150-300), their weights and the parameters of the Bernoullis and Gaussians.

In order to predict a whole protein structure, the amino acid sequence is broken into overlapping fragments using a sliding window. The overlapping regions are then modelled in k mixture distributions for fragments of length k. This introduces local biases to the structural preferences of the fragments. Combining the mixture models is not straight forward. A single site feels the influence of several distributions, which may not entirely agree with each other.

With this scoring framework we are able to generate protein structure samples in four steps. First, the class weights are calculated from the whole amino acid sequence. Then, the conformational space is narrowed down by iteratively updating the class weights of overlapping fragments by decreasing a temperature-like parameter, similar to simulated annealing. The local preferences are propagated, as the positions within a fragment are correlated. When the system cools down the consistent classes are favoured. From the conditional class weights sample structures can be generated. As a final step, steric clashes are removed and the models are collapsed to a realistic gyration radius by resampling random stretches in a naive greedy way. Unlike standard SCMF, the method works without assuming the Boltzmann relation at any stage.

In order to turn this into a prediction program, a huge number of samples is generated and ranked with the product over all fragment probabilities, i.e. the sum of the dihedral angle probabilities weighted by the conditional class weights.

The methods have been tested in an international comparison (CASP8) which takes place every two years. We have participated with four fully automatic servers[3], which turned out to give entertaining answers. Most of the time secondary structure elements were recognised correctly but somewhat misfolded, because of the weaknesses in our long range terms.

Finally, we have used our experience for the inverse problem, amino acid sequence optimisation. The sequence space is vast compared to the structure space. That means, for a given structure one can find many sequences that fold to it, whereas the opposite is seldom true. Therefore we are confident of finding putative sequences with our sampling approach. Following a similar path as the structure sampling/prediction method, our method is able to produce promising sequences. Our solutions seem to be more repetitive than natural proteins, but this may not be a problem. We also have
started to force the samples towards protein-like sequences by adding extra terms to the scoring function.

Preliminary results provide evidence that a combination with Monte-Carlo methods improves the situation, as these methods use importance sampling and therefore generate samples closer to the optimum of the scoring function. We want to include solvation terms into the scoring functions and eventually use these terms to guide the production of compact structures.

Furthermore, we want to improve the classification model by using circular distributions for modelling dihedral angles as has been done recently[4]. This would lead to a relatively elegant procedure as angles could be modelled more accurately. We want also to compare our scoring function to well established scoring functions in more detail.

Sometimes, in structural biology only parts of a protein structure could be solved. We would like to extend and apply our methods to the loop closing problem, when parts of a structure are not experimentally determined. We expect the problem to be sufficiently constrained to suit our sampling approach.

REFERENCES:
3. CASP Server URLs:
   http://cardigan.zbh.uni-hamburg.de/~mahsch/schenk/
   http://cardigan.zbh.uni-hamburg.de/~mahsch/mahmood/
   http://cardigan.zbh.uni-hamburg.de/~mahsch/qa-ms/
   http://www.zbh.uni-hamburg.de/wurst/

52. A new protein binding pocket similarity measure based on comparison of 3D atom clouds: application to ligand prediction.

Brice hoffmann, Mikhail Zaslavskiy, Jean-Philippe Vert and Véronique Stoven, Mines ParisTech, France.

We explore a new approach for ligand prediction in which binding pockets are represented by atom clouds. Pockets are aligned in 3D space with further use of convolution kernels between clouds of points. Performance of the new method for ligand prediction is compared to those of other available measures and to docking programs.

One of the main goals of structural biology is to predict, from the 3D fold of a protein, its interacting partners, which in turn is related to its molecular function. However, understanding this structure-function relationship is still today an open question, and no reliable tool is available to permit such a prediction. Current efforts concentrate on local 3D approaches, focusing on identification and comparison of binding pockets, in order to predict the natural ligand for a protein, with the underlying idea that proteins sharing similar binding sites are expected to bind similar ligands. The same strategy also applies to the problem of identifying new drug precursors for a therapeutic target protein.

We explore the potential of a new approach called sup-CK in which binding pockets are represented by clouds of atoms in 3D space potentially baring additional labels such as partial charge or atom type. The new similarity measure is based on the alignment of protein pockets with further use of convolution kernel between 3D point clouds. We study how the proposed method may be used to predict a ligand for a given pocket by comparing it to a set of pockets with known ligand. We also consider simple methods based on the comparison of simple binding pockets characteristics. These methods represent each pocket by an ellipsoid constructed on the basis of pocket principal axis. Since volume information was found to be important by Kahraman et al. (2007), we also test a linear combination of the sup-CK and volume-based methods, called sup-CK-Vol, where the coefficient of linear combination is learned as other model parameters in the double cross validation scheme. This linear combination takes advantage of the Vol method to separate very different pockets like PO4 and NAD, and of the sup-CK algorithm to allow finer discrimination.

The comparison of 3D binding pockets is an active field of research, and during the last decade, many new methods were proposed. Morris et al. (2005) proposed to model pockets by star-shapes built using the SURFNET program. Kahraman et al. (2007) presented three different variants of SHD, using only the shapes of binding pockets, the sizes of the binding pockets (keeping only the zero-th order in the spherical harmonics expansion), and their combination. Many binding pockets similarity measures are based on pocket alignment with further counting of overlapping atoms. In particular this kind of approach is used in the Poisson index model (Davies et al., 2007). Shulman-Peleg et al. (2008) represent pockets by pseudo-atoms labeled with physico-chemical properties. Pockets are aligned using a geometric hashing technique. This algorithm was mainly designed for multiple alignment of binding sites, but it may be used for pairwise alignment of pockets, as was performed in this study.
In our experiments, we extracted pockets on the basis of known protein-ligand crystal structures as it was done by Kahraman et al. (2007). In cases where the binding site is unknown, various programs have been developed to locate depressions on protein surfaces and could be used to identify putative binding sites (Glaser et al. 2006).

We consider several benchmark datasets. The first one comprises the crystal structures of 100 proteins in complex with one of ten ligands (AMP, ATP, PO4, GLC, FAD, HEM, FMN, EST, AND, NAD) and was proposed by Kahraman et al. (2007). We built an extended version of the Kahraman dataset in which we added protein structures in complex with one of the same ten ligands, leading to a total of 972 crystal structures. The added proteins present pairwise sequence identities less or equal to 30%, to avoid potential bias by inclusion of close homologs. The Kaharaman dataset comprises ligands of very different sizes and chemical natures. However, the real challenge is to test methods on pockets that bind ligands of similar size. Therefore, we created a third dataset comprising 100 structures of proteins in complex with ten ligands of similar size (ten pockets per ligand).

An important question is the evaluation of pocket similarity measures. We discuss two criteria to compare the quality of similarity measures on the basis of their ability to detect pockets binding the same ligand: area under ROC curve (AUC) and classification based scores. We compare our method with some existing state of the art algorithms on different benchmark datasets. Since we evaluate methods for binding pocket comparison according to their ability to predict ligands, we also report the performance of docking methods, on the same benchmark datasets. Finally, we also discuss possible extensions of the proposed method to other applications such as protein function prediction or ligand comparison. The Sup-CK method is available at: http://cbio.ensmp.fr/paris


53. Multiple Alignments and Phylogenies of Protein Structures.

Thomas Margraf and Andrew Torda. University of Hamburg Centre for Bioinformatics, Germany.

HANSWURST is a novel tool for fast multiple protein structure alignment. The algorithm quickly produces high quality alignments and can be used to align hundreds of proteins. These properties make it well suited for computing structure based phylogenies of large families of proteins such as the kinases.

The significance of matches in pairwise alignments can be difficult to judge against the background noise of random matches. In multiple alignments however, random matches across a reasonable number of structures are so improbable that there is little room for doubts about their significance. This is doubly true for multiple structure alignments which really begin to shine when the relationships between proteins become so remote that sequence methods start to break down. HANSWURST is built on the assumption that local interactions between atoms are the most important factor in determining the overall structure of a protein. Therefore, long stretches of high local similarity should also lead to high global similarity. From this reasoning follows that HANSWURST’s aim is not to produce alignments with optimal global similarity scores such as RMSD. Instead, good global scores are considered to be a property which emerges from local similarity. This is almost the exact opposite of the ideas behind some other traditional multiple structure alignment methods which sacrifice sensitivity for lower RMSD scores. Traditionally, molecular systematics has relied on sequence information in order to create taxonomies of organisms. Additionally, more and more structures of homologous proteins are becoming available. Since structure is generally more conserved than sequence, there is great potential to improve existing phylogenies by deriving evolutionary relationships from structure information.

The basis of the alignment method is a bayesian classification of protein structure fragments using the AutoClass program. Based on the class descriptions in such a classification, we can calculate the probability of a given protein fragment being in a certain class. The set of all class membership probabilities for a given protein fragment can be represented as a probability vector. The dot product of two such vectors can be used as a similarity measure between two peptide fragments. This score can then be used instead of a substitution matrix in standard sequence alignment methods. The resulting pairwise alignments of all vs. all structures one wishes to
align are then used to fill a distance matrix. On the basis of this matrix, various clustering algorithms can be used to construct a guide tree. Currently, the best such algorithm is derived from the UPGMA method and uses alignments of average probability vectors to estimate the distances between internal nodes of the guide tree. Such consensus probability vectors are computed by averaging the class membership probabilities of each fragment in a given column in the alignment. Gaps have no class memberships and thus do not contribute to the average. This concept allows each node in the guide tree to have a set of probability vectors associated with it which represent the average class memberships of that node's descendants. Since all the information required to compute a pairwise alignment is available for any cluster of structures, distances between clusters can be calculated by aligning their associated probability vectors. This removes the need to estimate distances during the construction of the guide tree and therefore improves its quality. The alignments of those average probability vectors are also used to merge the pairwise alignments according to the guide tree.

As a demonstration of our method's capabilities, we have built alignments and phylogenies of sets of up to 900 kinase structures. Computing this alignment took just over 8 hours of CPU time. Our guide trees represent purely structure-based phylogenies, but derived multiple sequence alignments are used to compare protein structures in a sequence-independent manner at a rate of thousands of proteins per second. It has been shown previously that the SPF approach has not yet been used to align and compare protein structures in a very efficient and completely sequence-independent manner. We present results which illustrate the evolutionary or functional relationships of proteins that have very low sequence similarity. However, whereas symbolic techniques such as the Smith-Waterman and Needleman-Wunsch algorithms for aligning protein sequences have become standard tools in bioinformatics, it remains an open question as to how best to align the 3D shapes of proteins (Sippel and Widerstein, 2008). Most existing structural alignment algorithms are based on comparisons of the Cα backbone traces or vectors formed by the Cα-Cβ atoms of each non-glycine amino acid, for example (Kolodny et al. 2005). However, these approaches are significantly more computationally expensive than the symbolic techniques because they typically entail the calculation of multiple least-squares rotation matrices, and this is very expensive in the context of dynamic programming alignment algorithms.

In a significant step towards performing protein structure comparisons more efficiently, Mak et al. (2007) and Sael et al. (2008) showed that the 3D shapes of large protein molecules could be compared and classified very rapidly using special 3D pose-invariant descriptors derived from spherical harmonic and Zernike polynomials (Novotny and Klein 2004). Hence this kind of approach offers the possibility of being able to search a 3D database of protein structures in a sequence-independent manner at a rate of thousands of proteins per second. It has been shown previously that the SPF approach provides fast way to perform protein-protein docking correlations (Ritchie and Kemp 2000, Ritchie et al. 2008). However, until now, the SPF approach has not yet been used to align and compare protein shapes. Conceptually, the use of Zernike descriptors has close parallels with the SPF representation, although both Mak et al. and Sael et al. did not exploit the special rotational properties of the spherical harmonic basis functions. Hence their approaches can identify similar protein shapes, but cannot align them.

Here, we demonstrate that 3D-Blast can quantitatively superpose and compare protein structures in a very efficient and completely sequence-independent manner. We present results which illustrate the accuracy with which protein shapes can be encoded and reconstructed using the SPF representation, and we compare our results with those of Mak et al. and Sael et al. on a selection of test cases. We also demonstrate the utility of our approach, by performing queries of single protein structures against the CATH database (Cuff et al., 2008).


55. Alpha-helical transmembrane protein fold prediction using residue contacts.

Timothy Nugent and David Jones, UCL, UK.

Despite significant efforts to predict TM protein topology, little attention has been directed toward developing a method to pack the helices together. We present a novel approach that uses predicted lipid exposure, residue contacts, sequence statistics and a force-directed algorithm to find the optimal helix packing arrangement.

We present a novel method to predict lipid exposure, residue-residue contacts, helix-helix contacts and finally the helical packing arrangement of TM proteins. Using molecular dynamics data to label residues exposed to lipid, we have trained and cross-validated a support vector machine (SVM) classifier to predict per residue lipid exposure with 70% accuracy. This information is combined with additional features, including sliding windows generated from PSI-BLAST profile data and a variety of sequence-based features, to train an additional SVM to predict residue-residue contacts. Combining these results with a priori topology information, we are able to predict helix-helix interaction with 69% accuracy using leave-one-out cross validation (LOOCV) on a test set of 74 protein chains. We then tested the ability of the method to discriminate native from decoy helical packing arrangement using a decoy set of 2811 structures. By comparing our predictions with the test set, we were able to identify the native packing arrangement with 72% accuracy. All these performance metrics represent significant improvements over existing methods. In order to visualise the global packing arrangement, we adopted a graph-based approach. By employing a force-directed algorithm, the method attempts to minimise edge crossing while maintaining uniform edge length, attributes common in native structures.

Our results demonstrate that the use of predicted lipid exposure data combined with evolutionary information and sequence-based statistics can be used to accurately predict the packing arrangement of TM proteins, and discriminate native from decoy packing arrangements. This method can be used to gain insights into TM protein folding and direct further experimental work.

Protein--protein interactions are challenging targets for modulation by small molecules. Here, we propose an approach that harnesses the increasing structural coverage of protein complexes to identify small molecules that may target protein interactions. Specifically, we identify ligand and protein binding sites that overlap upon alignment of homologous proteins. Of the 2,619 protein structure families observed to bind proteins, 1,028 also bind small molecules (250-1000 Da), and 197 exhibit a statistically significant (p < 0.01) overlap between ligand and protein binding positions. These ‘bi-functional positions’, which bind both ligands and proteins, are particularly enriched in tyrosine and tryptophan residues, similar to ‘energetic hotspots’ described previously, and are significantly less conserved than mono-functional and solvent exposed positions.

Homology transfer identifies ligands whose binding sites overlap at least 20% of the protein interface for 35% of domain--domain and 45% of domain--peptide mediated interactions. The analysis recovered known small-molecule modulators of protein interactions as well as predicted new interaction targets based on the sequence similarity of ligand binding sites. We illustrate the predictive utility of the method by suggesting structural mechanisms for the effects of sanglifehrin A on HIV virion production, bepridil on the cellular entry of anthrax edema factor, and fusicoccin on vertebrate developmental pathways. The results, available at http://pibase.janelia.org/ligands, represent a comprehensive collection of structurally-characterized modulators of protein interactions, and suggest that homologous structures are a useful resource for the rational design of interaction modulators.

57. Structural Signature of Antibiotic Binding Sites on the Ribosome.

Hilda David-Eden and Yael Mandel-Gutfreund, Faculty of Biology, Technion-Israel Institute of Technology, Haifa Israel

The ribosome is a complex molecular machine which offers many potential sites for functional interference. Here we have analyzed the structural and evolutionary properties of known antibiotic binding sites on the ribosome, identified by crystallography, and defined an RNA signature of ‘druggable’ pockets on the ribosome.
data set for understanding the unique features of drug binding pockets on the ribosome. In this work we have analyzed the structural [1, 2] and evolutionary properties of 33 antibiotic binding sites on the ribosome, identified by crystallography. We compared these sites to putative small molecule-binding pockets present in the small and large ribosomal subunits [3-5]. On the basis of this analysis we defined properties of the known drug binding sites which constitute an RNA signature of a ‘duggable’ pocket in the ribosome. The most noticeable property that defines true drug-binding sites was the prevalence of non-paired bases. In addition, in these sites we observed a strong bias to the unusual syn conformations of the RNA bases and the C2’ endo sugar pucker. We propose that albeit the different geometric and chemical properties of diverse antibiotics, their binding site tend to have common attributes, possibly reflecting the potency of the pocket for binding small organic molecules. The finding properties can be used to identify new ribosomal sites which could be targeted by small molecular weight ligands, including new antibiotics.

Figure: Overlapping between computed and observed binding pocket of erythromycin. The rRNA molecular surface is presented in grey; erythromycin is presented as blue sticks. The computed binding site which shows the highest overlapping is colored yellow. The pocket was calculated on the crystal structure of the unbound H. marismortui, (PDB IDs 1IJ2), and was superimposed on the bound structure (PDB ID 1YI2).

References

58. A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation.


We present a method to derive multi-body contact potentials, based on the arrangement of circles on sphere. Using this geometric construction on coarse-grained protein models, we show that we can build various knowledge-based potentials, encoding up to 5-body contacts, able to distinguish native structures from decoys.

Knowledge-based protein folding potentials have proven successful in protein folding. Based on statistics on interatomic distances, they generally encode pairwise contact information. In this study we present a method that derives multi-body contact potentials from measurements of surface areas using coarse-grained protein models. The measurements are made using a geometric construction: the arrangement of circles on a sphere. This construction allows the measure of residue covering areas which are used as parameters to build functions able to distinguish native structures from decoys. These functions, encoding up to 5-body contacts are evaluated on a reference set of 66 structures and its 45000 decoys, and also on the lattice ssfit set from the decoys’R us database. We show that the most relevant information for discrimination resides in 2- and 3-body contacts. These potentials could lead to different types of structure refinement techniques that use multi-body interactions.


The Electron Microscopy Data Bank is the international repository for high-resolution electron microscopy density maps, containing molecular structures as well as an increasing number of tomograms in the cellular context. Future development for this resource include supporting more complex imaging contexts, fitting and segmentation interfaces, and improvements in quality assessment.

The Electron Microscopy Data Bank is the international repository for high-resolution three-dimensional electron microscopy images of biological specimens. Established in 2002 at the European Bioinformatics Institute (EBI) in Cambridge, UK, it has become a standard resource for structural biology and is now operated jointly between the EBI and the Research Collaboratory for Structural
Biology (RCSB) at Rutgers University in New Jersey.

Most of its more than 500 entries are high-resolution three-dimensional density maps of macromolecules, based on computational averaging (‘single-particle’ method) of many individual lower-resolution images. This method can reach resolutions of 4 Angstrom, rivaling X-ray diffraction and providing a basis for solving atomic structures of large protein complexes. An increasing number of entries are three-dimensional tomographic maps, reconstructed from tilt series showing subcellular structures or individual macromolecules. In the future, cryo-electron tomography is expected to provide a bridge between structural, cellular, and systems biology in trying to understand macromolecular complexes as molecular machines in a cellular context.

To adapt to these new technologies, we are developing tools to represent entries that contain “rich content”, i.e. not only a three-dimensional map, but also atomic structure coordinates, averaged submaps for resolution improvement, ensembles of maps resulting from classification, and markup information such as segmentation masks. A standardized flexible representation based on content graphs is the basis for integration of computational analysis and visualization tools. AJAX technologies for dynamic web interfaces can provide users with simple and efficient submission, retrieval, and visualization tools, while comprehensive standardization efforts enable interoperability with the EM software toolchain.

For computational structural biology, the EM Databank will offer programmatic access to the maps as well as annotation interfaces to support segmentation and markup of structures as well as fitting of PDB coordinates into maps. With increasing resolution and complexity of the deposited data, new challenges will arise to design reliable algorithms in these areas.

A topic of increasing importance is quality assessment, which in the case of EM maps was always hampered by the large data sets of original images required for these. In cooperation with high-performance computing centers, we are developing new tools for the management of such data sets from the users workstation to the central repository and back. Integration with workflow tools for computational analysis of EM images and standardization of its representation supports data harvesting and direct deposition of complex workflows into the database, while standardized quality assessment tools are developed to operate on these data and provide user feedback.

Finally, quality assessment of three-dimensional density maps poses new computational and algorithmic challenges. Currently, very little analysis is expended to evaluate the quality of the resulting reconstructions. Statistical methods based on resampling not only require large amounts of computing time, but also a better characterization of the statistical probabilities of the reconstruction algorithms. Better quality measures for flexible fitting of PDB structures have to be developed that allow to evaluate reliably margins of errors in fitting.

With increasing amounts of data, both in increasing resolution and in covering the cellular context, electron microscopy images pose new challenges to computational structural biology. EMDB tries to provide a data repository and interface for solving these.

60. Automated Ligand-based Active Site Alignment.

Abraham Heifets and Ryan Lilien, University of Toronto, Canada.

We present LigAlign, a new system for ligand-based active site alignment. LigAlign automatically computes atom correspondences and aligns different conformations of flexible ligands by detecting rotatable hinge regions. LigAlign is an extension to the PyMOL molecular visualization tool and is freely available from http://compbio.cs.toronto.edu/ligalign

Protein—Ligand binding and subsequent enzyme function is mediated by a complex set of molecular interactions centered at the protein’s active site. Structural biologists aim to understand these interactions to gain insight into protein function, suggest opportunities for activity regulation and modulation, and elucidate
evolutionary history.

One promising technique for analyzing important features of protein-ligand interfaces is the characterization and identification of a set of physiochemical properties specific to the binding of various ligand types. An investigation of such physiochemical properties will be presented as part of the main conference program of ISMB 2009 [1]. Typically, common feature identification relies on the analysis of multiple aligned protein-ligand complexes. By aligning each structure, patterns may emerge. Therefore the approach has two parts: the generation of an informative structural alignment followed by the detection of physiochemical patterns within the aligned structures. We address both parts in our work.

The structural alignment can be protein- or ligand-based. In protein-based alignment, a transformation is computed to superimpose a selection of residues lining the active site. In contrast, ligand-based alignment computes a transformation to superimpose the corresponding atoms of the target ligands. Although the required atom-based correspondence is trivial to compute when the ligands are identical; building a correspondence for similar yet non-identical ligands (such as those generated for SAR or high-throughput studies) is more complex. Nevertheless, ligand-based alignment is more likely to superimpose structurally relevant regions of the complex, particularly in case of functionally homologous proteins which bind similar ligands using different active site residues.

Regardless of the alignment type, it is important to model the molecular flexibility of the ligand. A common way to model this flexibility is by the introduction of internal torsion angles or hinges which divide the ligand into rigid fragments. While a flexible ligand may bind two different proteins in two different conformations, the local sub-cavity surrounding each corresponding fragment is likely to be similar. In the absence of a flexibility model, the desired interaction patterns are unlikely to coincide in the selected coordinate frame and pattern detection is compromised [2][3][4].

We present a new flexible ligand-based active site alignment algorithm, LigAlign. Given a set of protein-ligand complexes, LigAlign automatically identifies an atom correspondence, detects flexible hinges, computes the fragment-based local alignments, and graphically displays the results. By automating the steps of a traditional ligand-based alignment algorithm, LigAlign reduces the amount of user intervention and widens the scope of analyzable ligands. Our software is implemented as an extension to the PyMOL molecular visualization tool [5], which permits easy integration with other structural studies.

The LigAlign algorithm can be divided into three major components: building an atom correspondence, ligand fragmentation, and cluster detection. For clarity, in the following, we describe aligning a single flexible query ligand against an anchor ligand. In practice, our algorithm can align any number of flexible ligands.

Atom Correspondence: For each atom in the query and anchor ligands, we first generate a fingerprint based on the structure of the atom's chemical properties and nearby molecular branching and connectivity. Using these fingerprints, we calculate a pairwise similarity score between every atom in the query ligand and every atom in the anchor. Finally, we generate a correspondence by solving a weighted bipartite matching problem, where the nodes are the atoms from the two ligands and the edge weights are computed from the similarity scores. This method allows an atom correspondence to be constructed even for distinct ligand analogues.

Ligand Fragmentation: Rigidly aligning flexible ligands may yield poor overlap due to the variability of ligand conformations in different protein complexes. Therefore, our algorithm fragments the query ligand by inserting up to a user-specified maximum number of rotatable hinges. Because the system lacks foreknowledge of the correct hinge placements, we perform a search of possible hinge placements to find the fragments which yield the smallest total RMSD when aligned independently to the anchor ligand. We use dynamic programming and branch-and-bound to prune the search space and find an exact solution quickly.

Cluster Detection: In cases where the superposition of multiple proteins reveals clusters of conserved amino acids, LigAlign presents these conserved protein structures for visual inspection via PyMOL. Clusters are defined to contain residues where, after alignment, corresponding residues in different proteins fall within a user-specified maximum distance. It follows from this definition that, for k proteins, cluster detection is an instance of the NP-hard k-partite-epsilon-pivot maximum cardinality matching in 3 dimensions problem [6]. Despite this worst-case complexity, we note that empirically, instances are quick to complete.

In this work, we present the results of our automatic ligand-based alignment method using several experimental test systems. When run on the systems introduced by Najmanovich et al. and Nebel [3][7], we generate interaction profiles consistent with those previously identified. We also present early results for systems with less conserved ligands, particularly flexible ligands in multiple conformations, such as FAD and NAD. In both cases, our PyMOL interface allows easy visualization. Source code freely available from http://compbio.cs.toronto.edu/ligalign under GNU LGPL, implemented in Python, and supported on Apple OS X, Linux and MS Windows.

FIGURE LEGEND:
Figure on the left shows a computed atom correspondence for imatinib.

REFERENCES:

A graphical model approach for predicting free energies of association for protein-protein interactions under backbone and side-chain flexibility.
We introduce GOBLIN, a method that predicts binding free energies of protein-protein interactions by approximating the partition function of the Boltzmann distribution. Using the ROSETTA force field to compute internal energies, GOBLIN achieved RMSEs of 2.05 kcal/mol with experimental DeltaDeltaG on a benchmark set, nearly 0.55 kcal/mol better than ROSETTA.

GOBLIN uses a sample of backbone configurations using ROSETTA’s backrub mode around the crystal backbone to model P(Xb=b) (backbone flexibility). Computing Zb for each backbone requires a summation over all possible rotameric states Xs that are consistent with b. To perform this exponentially large summation, GOBLIN exploits independencies present in the distribution of side-chain configurations Xs conditioned on specific backbone traces b, P(Xs|Xb = b) that exist due to the rapidly-decaying nature of the physical energies, to compactly encode it as a Markov Random Field (MRF). The attached figure shows a part of the MRF encoding the conditional distribution P(Xs|Xb=b) for Chymotrypsin complexed with the third domain of turkey ovomucoid. Solid lines refer to intra-molecular interactions, and dashed lines refer to inter-molecular interactions. The pifs represent matrices containing the Boltzmann factors between all pairs of side-chain configurations for residues sharing an edge that are stored along with the graph structure in the MRF.

Using ROSETTA’s force-field to compute the internal energy of each configuration, GOBLIN then uses an approximate inference algorithm called Belief Propagation (BP) [3] on this MRF to compute an approximation to Zb. Significantly, it has been shown that using BP computes the Bethe approximation of the free energy [4]. Thus, there is reason to believe that these approximations are also physically valid.

We studied the efficacy of our approach on a database of over 700 single-point mutants from eight large and well studied complexes for which experimentally measured changes in free energy are available. Our results [1] show that GOBLIN is accurate, with root mean squared errors (RMSE) of 2.05 kcal/mol relative to experimental values. Significantly, our method outperforms the well-known program ROSETTA in terms of accuracy by 0.55 kcal/mol. This result is especially interesting because we implemented ROSETTA’s own force field in order to compute internal energies. That is, our improved accuracy can be attributed to our approach to incorporating protein flexibility.

We also compared the accuracy of various settings of GOBLIN: side-chain flexibility only; side-chain flexibility with parameters optimized for a specific dataset (GOBLIN-spec); backbone and side-chain flexibility with parameters optimized for a specific dataset. This ‘-spec’ settings use a novel algorithm that we have developed to learn force-field parameters to minimize the mean square error (MSE) in predictions over a training set from the same complex. Our results show that GOBLIN-spec RMSEs are lower than GOBLIN by 0.26 kcal/mol on average. A more detailed presentation of this work is available in [1] while a more rigorous treatment of the statistics is available in [5].

Fast and accurate free energy calculations are essential to a number of significant tasks within Computational Structural Biology, including structure-based protein and drug design. Our probabilistic graphical model-based approach to all-atom free energy calculations strikes a balance between the rigor of physical methods (i.e., molecular dynamics based free energy calculations) and the speed of statistical methods. Our method is physically rigorous in that (i) it uses all-atom force fields when computing internal energies, and (ii) it computes a rigorous approximation of the true partition function of the system. At the same time, our method is competitive with statistical methods, in terms of speed,
typically requiring less than 5 minutes per calculation.

References


We present a general, high-resolution modeling protocol for peptide-protein interactions. FlexPepDock has been adapted from the RosettaDock protein docking protocol to refine approximate structures (3.0 Angstrom peptide atom RMSD) to high-resolution structures (<1.5 Angstrom RMSD) - detailed enough for structure-based design of inhibitory peptides and other drug molecules.

A wide range of regulatory processes in the cell is mediated by flexible peptides that fold upon binding onto a globular protein, involving substantial conformational changes of the backbone and side-chains. Accurate, atomic-resolution models of peptide-protein binding interactions are essential for studying this binding mechanism in detail. They can be used for the design of inhibitory peptides and peptide-derived drug molecules, just as experimentally determined crystal structures of peptide-protein interactions. While modeling protocols have been developed for specific peptide-protein interactions (e.g. PDZ-DocScheme for peptide-PDZ domain; Niv et al. 2005 JACS 127:14072), no broadly applicable protocol for the high-resolution, atom-level modeling of a range of different peptide-protein interactions has yet been reported.

We present FlexPepDock, a general protocol for the high-resolution modeling of peptide-protein interactions. FlexPepDock is derived from the high-resolution RosettaDock docking protocol (Gray et al. 2003 JMB 331:281) and is designed to allow significant changes in the peptide backbone. In each of the Monte Carlo minimization cycles, we alternately optimize either the peptide backbone or the rigid body orientation. Side-chain rotamers are optimized on-the-fly after each perturbation. We create k candidate decoy structures (typically 100-1000), rank them based on the Rosetta full-atom energy score, and select the top scoring decoys as high-resolution modeling predictions (Figure 1, inset).

A non-redundant subset of 15 peptide-protein complexes was extracted from the PeptiDB database of peptide-protein complex structures (London et al.; under preparation). To assess performance of FlexPepDock, a series of tests of increasing difficulty levels was devised. We first verified that the native structure resides in an energy minimum of the Rosetta energy function: gradient-based energy minimization kept 12/15 structures within 0.75 Angstrom of the Rosetta energy minimum of the RosettaDock docking protocol (Gray et al. 2003 JMB 331:281) and is designed to allow significant changes in the peptide backbone. In each of the Monte Carlo minimization cycles, we alternately optimize either the peptide backbone or the rigid body orientation. Side-chain rotamers are optimized on-the-fly after each perturbation. We create k candidate decoy structures (typically 100-1000), rank them based on the Rosetta full-atom energy score, and select the top scoring decoys as high-resolution modeling predictions (Figure 1, inset).

We then investigated two central features of the protocol, namely model accuracy and applicability range. Our aim was to create high-resolution models of atomic accuracy that will enhance the contribution to structure based computational analysis and manipulation of an interaction. In addition, the protocol should be applicable to a wide range of different starting structures: the larger the basin of attraction, the more structures can successfully be refined to high resolution, and the more systems will be amenable to high-resolution modeling by our protocol. For this purpose, we examined the performance of our protocol on a dataset of peptide-protein interactions, by systematically perturbing the native peptide backbone angles and rigid body orientations of initial structures. The results indicate that for starting structures within 3A peptide backbone RMSD (pepbRMSD) from the native complex, FlexPepDock samples at least one near-native decoy (within 1A pepbRMSD) in 80% of the cases, and in 40% (60%) it is ranked as the top (top 5) scoring solution(s). Figure 1, left panel, shows an example of a 3.5A starting pepbRMSD structure refined to 0.5A. In a number of cases the correct solution can be detected even when starting from farer away: Figure 1, right panel, shows an example of a 5.9A starting pepbRMSD structure refined to 1.2A. These results did not change dramatically when the experiment was repeated using the free protein structure: For starting structures within the effective range of the protocol (2.5-4A pepbRMSD), the sampling of high-resolution decoys (<2.0A pepbRMSD) is not
strongly affected: 91% successful sampling for the bound set vs. 88% for the free set. The ranking of these solutions is also similar: 89% vs. 83% ranked a near native model within the top 5 solutions for the bound and free protein, respectively. We had previously observed in the PeptiDB dataset that proteins do not significantly change conformation upon binding a peptide, and thus it does not come as a surprise that our protocol performs well also on free protein structures.

In order to simulate a real-world scenario where only sparse information about the peptide-binding site is known, we created peptides with an approximately correctly positioned anchor residue and an extended backbone. Starting from this conformation, FlexPepDock was able to select a model within 1.5A pepbbRMSD in 9/15 cases (no helical peptides were included in that dataset).

Another real-world scenario is the modeling of additional peptides based on a known peptide-protein complex structure, e.g. for example the binding of C-terminal peptide motifs to PDZ domains. In 6/7 cases, the top-scoring model was within 2.2A pepRMSD from the native structure. In 5/7 cases, our generic protocol FlexPepDock outperformed PDZ-DocScheme, and in the remaining two cases, comparable models were created. We note that FlexPepDock does a significantly better job in selecting good models by score.

In summary, we have described the development and assessment of FlexPepDock. We show that (1) highly accurate models (within 1-2A pepRMSD) are obtained from coarse models (with <3-4 starting pepRMSD), and that (2) the modeling accuracy is robust to using either the bound or the free backbone of the target protein. We examine the utility of our protocol over several realistic scenarios, including: (1) Docking from an extended backbone conformation, using only partial information of the binding site location, and (2) cross-docking using complexes of alternate peptides with the same receptor, where we show improvement of our general protocol over a protocol developed for a specific system.

Our current protocol provides the final and crucial step in which coarse computational predictions are refined into high-resolution models, at a level of accuracy comparable to crystallographic data (see Figure 1). Currently, efforts are put on extending the range of the protocol by including more aggressive sampling techniques used in the Rosetta ab initio protocol. In parallel, we are optimizing a complementary, low-resolution protocol for the definition of the approximate location of the peptide-binding site on proteins. Successful protein-protein docking protocols consist of a combination of a coarse-grained global search that detects starting structures for a subsequent high-resolution local refinement step (e.g. by RosettaDock). An effective combination of a low-resolution protocol with FlexPepDock may thus well provide a full ab initio protocol for the reliable creation of high-resolution peptide-protein complex structures.

63. How good can template-based modelling be?


Template-based models may be improved by over 10% GDT-HA via modifying the position of the template fragments. The extent of improvement is strongly predicted by sequence identity, helical fraction of structure, the number of fragments in the alignment, and the mean fragment length.

In summary, we have described the development and assessment of FlexPepDock. We show that (1) highly accurate models (within 1-2A pepRMSD) are obtained from coarse models (with <3-4 starting pepRMSD), and that (2) the modeling accuracy is robust to using either the bound or the free backbone of the target protein. We examine the utility of our protocol over several realistic scenarios, including: (1) Docking from an extended backbone conformation, using only partial information of the binding site location, and (2) cross-docking using complexes of alternate peptides with the same receptor, where we show improvement of our general protocol over a protocol developed for a specific system.

Our current protocol provides the final and crucial step in which coarse computational predictions are refined into high-resolution models, at a level of accuracy comparable to crystallographic data (see Figure 1). Currently, efforts are put on extending the range of the protocol by including more aggressive sampling techniques used in the Rosetta ab initio protocol. In parallel, we are optimizing a complementary, low-resolution protocol for the definition of the approximate location of the peptide-binding site on proteins. Successful protein-protein docking protocols consist of a combination of a coarse-grained global search that detects starting structures for a subsequent high-resolution local refinement step (e.g. by RosettaDock). An effective combination of a low-resolution protocol with FlexPepDock may thus well provide a full ab initio protocol for the reliable creation of high-resolution peptide-protein complex structures.
for loop modelling.

These results demonstrate that substantial improvement could be made on many templates if the conserved fragments were to be optimally positioned, and provides an accurate basis for identifying those templates for which modifying fragment positions may yield substantial improvements.

64. Ligand-based and structure-based virtual screening to identify potential inhibitors of drug-efflux transporters.

Sergio Mares-Samano, Raj Badhan and Jeffrey Penny. The University of Manchester, UK.

Pharmacophore structures of P-glycoprotein (P-gp/ABCB1) and Breast Cancer Resistance Protein (BCRP/ABCG2) modulators were constructed and used to screen the ZINC database. Hits were docked into homology models of the Nucleotide-Binding Domains (NBDs) of both transporters. Based on binding mode and predicted affinity, potential drug-efflux transporter inhibitors were identified.

P-glycoprotein and Breast Cancer Resistance Protein are ATP-dependent drug-efflux transporters that pump drugs out of cells [1]. Both proteins are expressed at low levels in many tissues but are highly expressed in intestinal and renal epithelial cells and endothelial cells of the blood-brain barrier [2]. This expression profile may influence the pharmacokinetic characteristics of drug substrates [3] and in vitro and in vivo studies have consistently shown that both proteins limit the intracellular accumulation of a broad range of drugs, including antivirals, antibiotics, calcium channel blockers and anti-cancer agents. Furthermore, overexpression of P-gp and BCRP is related to the multidrug resistance (MDR) phenotype, a phenomenon whereby cancer cells acquire resistance to a broad spectrum of structurally unrelated drugs [4]. This phenomenon has also been observed in tuberculosis, malaria and bacterial and HIV infections [5]. As a consequence, efficacy of drug treatment with P-gp and BCRP substrates is frequently poor and could lead to fatality. Due to such broad substrate recognition, inhibition of P-gp and BCRP may potentially improve the efficacy of a range of disease treatments.

ATP-binding and subsequent hydrolysis within nucleotide-binding domains (NBDs) of both transporters are essential processes whereby the energy released is utilised to trigger transport activity. Therefore, disrupting these processes is an extremely attractive approach to inhibiting drug efflux activity.

High-throughput screening of chemical libraries is a well-established method employed to identify new lead compounds [6]. However, available databases are constantly increasing in size and in vitro screening of compounds is extremely expensive and time consuming. Virtual screening of chemical libraries, a time- and cost-effective approach, is hence being increasingly used to streamline the drug discovery process [7].

Here we describe the implementation of a strategy that combines ligand-based and structure-based virtual screening techniques to identify potential P-glycoprotein and BCRP inhibitors targeted to the nucleotide-binding domains. The commercially available subset of the ZINC database [8], comprising approximately nine million compounds, was the chemical library analysed.

Ligand-based (pharmacophore-based) virtual screening is aimed at identifying molecules, with physicochemical similarities to known ligands, that are likely to interact with the target. Pharmacophore generation requires a representative set of known ligand structures with molecular descriptors [9]. Consequently, construction of pharmacophore models was based on chemical and topological fingerprints and also molecular descriptors of a series of compounds, in particular flavonoid derivatives, which are reported to interact with NBDs in vitro [10]. Analyses of pharmacophore features revealed key physical, chemical and spatial constraints that are determinants of potential inhibitory activity. These pharmacophoric features were used as an in silico filter to screen the ZINC database. Approximately 8500 hits were obtained. Pharmacophore generation and pharmacophore-based screening were carried out using Instant JChem 2.4.3.1 [11] and Screen 5.2.0 suites of programs [12].

Structure-based screening (in silico docking), which requires the 3D structure of the protein receptor, is aimed at predicting the affinity between ligands and a target protein [9]. Previously generated structural homology models of P-gp and BCRP NBDs [13], were therefore used to dock the hits using DOVIS 2.0 [12]. Based on predicted affinity and binding orientation within the ATP-binding pocket of NBDs, compounds possessing predicted high binding affinity were identified. Future in vitro studies will be carried out to examine the potential ability of hits to inhibit drug transport activity mediated by P-gp and BCRP. Molecular dynamics will be carried out to gain a better understanding of the mechanisms of ligand-receptor interaction.

References
65. Simulation of induced structural transitions in an isolated Kv1.2/2.1 voltage-sensor important for the gating mechanism.

Christine S. Schwaiger, Pär Bjelkmar and Erik Lindahl. Center for Biomembrane Research, Department of Biophysics and Biochemistry, Stockholm University, Sweden.

The Kv1.2/2.1 voltage-sensor movement is studied through an AFM-like technique. The voltage-sensor is pulled in two conformations (alpha/310), over the hydrophobic core proposed to be the gating free energy barrier, and the work distribution difference is compared. Additionally, residues responsible for the barrier are determined by utilizing core region mutations.

Voltage-gated channels open in response to voltage across the membrane, and control functions such as nerve signals, heart beat, and the way an egg closes after being fertilized by the first sperm. The highly, positive, charged S4 helix of their voltage-sensor domain (VSD) is supposed to move in response to an electric field, leading to a conformational change responsible for the gating mechanism. However, how this translation is accomplished still remains unclear and just hypotheses are available. This is especially due to the lack of crystal structures of the closed and intermediate states. As it has been shown that the VSD is functional even as a monomer (1, 2), we use the isolated structure of Kv1.2/2.1 as a model system to study how voltage activation works.

In ion channel crystal structures the S4 segment is mostly alpha helical. The Kv1.2/2.1 paddle-chimera channel (3) is the first structure identified where just the first part of S4 forms an alpha helix whereas the voltage sensor’s intracellular end adopts a ten residue long 310 helix conformation. This has risen the hypothesis that a secondary structure alteration could be critical in the gating mechanism (3). The 310 helix’s relevance for the translation is also supported by a recent experimental study (4) suggesting that between activated and the active state the S4 segment is present as a 310 helix whereas it converts into an alpha helix when moving between the activated and the inactivated state. This hypothesis has been further strengthened through over 1 microsecond molecular dynamics simulation (5), showing that structural rearrangements such as sensor rotation and an increase in 310 helix are caused by applied potential. This structural transition from alpha to 310 helix might be one of the key differences between the open-inactivated state of the crystal structure and resting and activated states where the channel is closed.

In this study we are studying the voltage sensor movement through umbrella sampling, a simulation technique similar to atomic force microscopy, using the GROMACS pull code (6). First, a transition to 310 helix is gradually induced by varying the backbone dihedral angles. Then, by pulling on the voltage-sensing Arginines, both the alpha and 310 S4 helix are dragged towards the intracellular side, over the hydrophobic core that has been proposed to be the free energy barrier of gating, and we compare the difference in work distributions. Additionally, we try to identify the residues responsible for the barrier by utilizing different mutation systems of the core region (e.g. F233, I177, I230) in combination with the pulling approach.

References:
Hidden Markov Models (HMMs) are frequently used to encode protein sequence information into stochastic models which subsequently are applied to identify related proteins in sequence databases. Profile HMMs are derived from multiple sequence alignments and thus heavily depend on the quality and information content of the latter. It has been shown that multiple structure comparison improves the performance of HMMs when the sequence similarity in a certain protein family is low [1].

Further improvement is expected when supplementary biological information such as strictly conserved sites can be considered more explicitly in the HMM. Working along these lines we developed methodologies and software tools which facilitate this flow of information by (i) the computation of reliable multiple structure alignments (MStAs) and (ii) providing the expert user a facility to fine tune the representation of the biological properties in the HMMs derived from the MStAs.

The resulting tools PSC++ and HMMModeler are implemented as an extension for the UCSF Chimera [2] molecular modelling system. As for the MStAs, PSC++ uses Ant Colony Optimization to build multiple structure alignments from alternative pairwise solutions [3]. HMMModeler [4] then provides for a GUI for editing the alignments and supports the user in adding knowledge to the model. This is done by column-wise fine tuning of predetermined breaking points (acceptance of insertions or deletions), degree of conservation, and specified amino acid sets. It allows for a precise definition of the skeleton of the HMM without the need for understanding the stochastic nature and numerous transition and emission probabilities of the model. Finally, HMMModeler generates a tailored 9-state profile HMM for the family under consideration. Scoring and alignment of single query sequences as well as automated database searches can be triggered. The approach was tested with SCOP families. Improvements compared to conventional unsupervised purely sequence based HMM methods are shown.

References:
The evaluation is based on comparing method results to curated reference alignments. The reference alignment test suite consists of two sets, RIPC and SISY-pairwise, for pairwise alignment accuracy [1] and one test set, SISY-multiple, for multiple alignment accuracy. All sets reflect difficult cases. The RIPC set was manually derived from SCOP[2] and corresponding literature references. The SISY set was derived from the SISYPHUS database, an assembly of alignments of proteins with non-trivial structural relationships[3]. To guarantee high quality and usability, the SISYPHUS alignments were processed by several filtering steps. Family members which cause ambiguities in the evaluation have been removed and the entries were adapted to the remediated PDB. Finally, all reference alignments were stored in a XML format suitable for pairwise and multiple structure alignments. All test cases were annotated for containing permutations, large InDels or conformational variations as well as possible caveats in the corresponding PDB files. All data are provided on our web pages at http://biwww.che.sbg.ac.at/RSA.

Several Python scripts were implemented to convert alignments from the raw methods output into the XML format. These parser scripts are available for the pairwise alignment methods CE, C-alpha match, DALI, FATCAT, MATRAS, SHEBA and TOPMATCH and the multiple alignment methods MAMMOTH-mult, MASS, Matt, MultiProt, MUSTANG and POSA. We further developed a USCF Chimera[4] extension which reads and visualizes the XML formatted alignments.

Using the XML formatted alignments, correctness of the method alignment compared to the reference alignments becomes straightforward. Results for the methods mentioned above will be presented.

References

68. Shedding light on transcription factor binding sites using molecular dynamics.


We describe a computational protocol for designing protein-DNA complexes, which have not been determined crystallographically. Using the MM-PBSA approach, we predict binding affinities between a WRKY transcription factor and different DNA sequences. The complexes which comprise known DNA binding sites were placed in first and fourth rank.

Introduction:
A key step in regulating gene expression is the sequence-specific binding of transcription factors (TFs) to their DNA recognition sites. TFs recognize and bind to short defined nucleotide sequence motifs (cis-elements), which are specific sites in the noncoding regulatory region in close proximity to a gene. WRKY proteins constitute a plant specific subfamily of the large FLYWICH-GCM1 super family of zinc finger TFs. They act as important regulators in developmental processes, pathogen defense, and in response to abiotic stresses. The WRKY protein family can be categorized into three distinct groups due to the overall structure of the proteins [1]. Nearly all WRKY proteins appear to have a stereotypic binding preference to a DNA element, termed the W-box: 5'- TTGAC-C/T  -3' [2]. Since no crystal structure of a WRKY-DNA complex is available in the PDB it would be highly beneficial to be able to design such a complex. Studying the binding interfaces of WRKY-DNA complexes can give a more complete understanding.
of specific binding and connections between the structure and function of WRKY proteins.

Materials and Methods:
One WRKY protein crystal structure (PDB id: 2ayd) [3] is available in the PDB, however, without DNA. This structure contains the C-terminal domain of Arabidopsis thaliana WRKY1 (AtWRKY1-C). We used the DALI server [4], a service for comparing protein structures in 3D, to find structures similar to AtWRKY1-C. The most similar protein structure is Drosophila GCM1 (PDB id: 1ohd) [5], a TF bound to DNA. Drosophila GCM1 obtains a comparatively high DALI Z-score of 6.2 and comprises a zinc finger motif like the WRKY proteins.

When the AtWRKY1-C protein structure is superimposed with the GCM1 protein structure the zinc finger motif and all secondary structure elements overlap perfectly. This yields a preliminary WRKY1-DNA complex. The 12 base pair (bp) DNA double helix of this preliminary WRKY1-DNA complex contains the consensus binding site of GCM1, which is completely different from the WRKY1 binding site. Therefore, a new DNA double helix was created using nucgen, a program provided by AMBER 10 [6]. The 32 bp long DNA sequence is identical to the parsley PR1-1 promoter containing one W-box and another known binding sequence similar to the W-box.

We let the WRKY1 TF slide along the PRI-1 promoter region in one bp steps while binding to the DNA major groove. For each step a protein-DNA complex is created, each with the WRKY1 protein and a 12 bp long DNA double helix. Since the orientation of the DNA is unknown both 5'-3' directions need to be considered, which yields 42 protein-DNA complexes.

Each WRKY-DNA complex is simulated in explicit solvent (TIP3P water model) at neutral pH using AMBER 10. The salt concentration is set to 0.2 M and each zinc ion is tetrahedrally coordinated by two water model) at neutral pH using AMBER 10. The salt concentration is set to 0.2 M and each zinc ion is tetrahedrally coordinated by two water residues. The production run of the MD simulation is 6 ns long. The AMBER 10 MM-PBSA script is used to calculate binding free energies of each WRKY-DNA complex. For these calculations the last 4 ns of the production run were used.

Results and Discussion:
We designed WRKY-DNA complexes and identified the energetically most favorable ones using molecular dynamic simulations with the MM-PBSA approach. The highest binding affinity is predicted for the sequence similar to the W-box: 5'-TTCAGC-3'. The experimentally well-studied W-box binding site, containing the DNA W-box sequence is identified with a high binding energy for those two complexes are probably caused by the high similarity of their binding sites to the experimentally known ones. However, the WRKY1-DNA complex containing the DNA W-box sequence is identified with a high binding free energy, which demonstrates the structural correctness of this protein-DNA complex. This WRKY1-DNA complex can serve as a basis for future experimental and computational studies on finding connections between the structure and function within the family of WRKY proteins.


69. Automated classification of enzymes using the Active Site Classification (ASC) method


Active Site Classification (ASC) is a method to predict the specificity within an enzyme family using sequence and structural information about the active site. We applied ASC to two enzyme families and could achieve improvements in classification accuracy and also interpretability of the models.

Enzymes that catalyse a similar general chemical reaction but use different substrates can often be grouped into families based on sequence similarity. Active Site Classification is a method to predict the specificity of an enzyme of such a family using sequence and structural information about the active site. Input is a set of sequences with annotated substrate specificity and a template structure for the enzyme family. First a multiple sequence alignment (MSA) of the input sequences is generated. Then the residues lining the active site are extracted from the template structure and the corresponding columns in the MSA are extracted to get an active site signature for each sequence. These signatures are then encoded into numerical feature vectors and a linear Support Vector Machine model is trained on this data. The model can be used to predict the specificity of test sequences from this enzyme family. Furthermore the learned model can interpreted to give insight into the mechanism of substrate specificity on a structural level. This is accomplished by sorting the model weights that are given to each feature by absolute magnitude and getting thereby a ranking of the importance of the active site residues. We applied the ASC pipeline to several enzyme classification tasks and analysed the residues found important in the structural context of the active site.
In the case of the classification task of creatine vs. arginine kinases we could build a model that achieved a crossvalidated classification accuracy of 100%. The baseline classifier, a one Nearest Neighbour (1-NN) classifier that emulates a BLAST-like prediction of function, achieved 97% accuracy. Although the 1-NN classifier can be considered quite competitive the ASC model is significantly better (t-test, p=1e-10) and it can furthermore be interpreted more easily than the instance-based 1-NN classifier. The three top ranked residues in the template structure were inspected in their structural context and one of these residues is believed to be a switch that influences whether creatine or arginine is used by enzymes of this family.

Similarly, for the decarboxylating dehydrogenase family an ASC model could be built that achieved an crossvalidated classification accuracy of 99.5%. The ASC model was significantly better (t-test, p=1.6e-6) than the baseline 1-NN classifier which achieved an accuracy of 97.5%. Interpretation of the learned model that discriminates between the isopropylmalate specificity and the isocitrate specificity yielded two residues that are important residues for determining substrate specificity.

In general, focusing the classification task on the active site improves accuracy significantly in the examples considered. The 1-NN classifier is competitive but not as easily interpretable as the ASC models.

70. TopDomain-web: interactive protein domain decomposition


TopDomain-web is an interactive interface for the domain decompositions used in the COPS protein classification. The service uses an efficient domain decomposition algorithm which recognizes single and multi-chain domains within complete biological units. The user can immediately visualize the decomposition and may interactively change the domain decomposition parameters.

We present an interactive web interface to visualize the domain decompositions used in COPS (Classification of Protein Structures [1,2,3]). The TopDomain-web service is a powerful tool to explore the domain structure of complete proteins and to analyse alternative solutions.

Protein domains are a widely used concept in protein biology and that concept has been around since the 1970s [4]. They are the basic units for classifying protein structures and annotating functions of proteins. The assignment of domains within a protein structure is ambiguous and there are many cases where alternative solutions may be acceptable.

The COPS classification uses the TopDomain program which implements an algorithm to assign structural protein domains based on the strength of interactions among amino acid residues (to be published). The algorithm is very efficient and enables the decomposition of entire biological units in real time. Nevertheless, it frequently happens that a domain decomposition is not unique. The web-interface is specifically designed to investigate alternative solutions and to study the domain decomposition problem in general. To this end the web interface provides access to a small set of parameters that can be used to investigate the stability of the domain decomposition of a particular protein. The TopDomain service is tightly integrated with the COPS system. In particular, domain decompositions are visualised instantly using Jmol and an interacting and sortable domain table with corresponding colours. Also, using the table's drag and drop functionality TopDomain-web is connected to the structure comparison program TopMatch-web [5] so that possible similarities among individual domains are recognized immediately. Moreover a right click triggers the qCOPS sequence search [6] using the selected domain as query. Each row in the grid can be dragged to the input box of TopDomain-web to start a new decomposition using the domains generated by TopDomain as queries. In this way the whole COPS system is accessible from the TopDomain service. It is also possible to upload files in the pdb format to investigate proteins that are not available in the PDB repository. The integration of TopDomain-web with the COPS system enables the user to investigate protein structures with an unprecedented efficiency. The TopDomain web-service is accessible at http://topdomain.services.came.sbg.ac.at

References:
[1] Stefan J. Suhrer, Markus Wiederstein, Markus Gruber and Manfred J. Sippl COPS - A Novel Workbench for Explorations in Fold Space Nucl. Acids Res. 2009 37(Web Server issue); accepted

71. COPS - A Novel Workbench for Explorations in Fold Space

Stefan J. Suhrer and Manfred J. Sippl. University of Salzburg, Salzburg, Austria.
The COPS web server is a next-generation web application that provides fast and intuitive access to the entire set of currently available protein structures. COPS organizes structural domains by quantified structural similarities which can be visualized immediately. Structural biologists can employ iCOPS to classify their own structures. Availability: http://cops.services.came.sbg.ac.at/.

Efficient access to the enormous body of information contained in the PDB (Protein Data Bank, (1)) requires that the protein structures are organized and classified according to a set of appropriate rules and principles. Moreover, the classification of protein structures implies several intricate tasks like the characterization of the classification unit or the definition of the properties to organize these units. But perhaps most challenging is the proper presentation of a classification scheme to enable users to easily access and browse the data and visualize structural similarities for the various tasks encountered in structural bioinformatics (8).

The COPS web server is a next generation web application that includes the Fold Space Navigator which interfaces the domain based Classification Of Protein Structures (COPS, (2)). COPS covers the complete repertoire of known protein structures and it is weekly updated with every PDB release. Pairwise structural similarities in COPS are encoded as quantified metric relationships and the resulting metrical structure is mapped to a hierarchical tree which is largely equivalent to the structure of a file browser (3). We take advantage of this analogy in the Fold Space Navigator to provide easy access to the domains in the classification.

Directory icons in the Fold Space Navigator group together domains that have a quantified degree of similarity to the representative domain of the respective folder. The similarity cutoffs of particular folders are defined by the hierarchical layers of COPS (2,4,5,7,8). For example, the domains within the folders on the lowest layer (Equivalent layer) have the highest structural similarities, and the folders on the highest layer (Distant layer), i.e. next to the root of the tree, contain pairs of structures that may have only weak structural similarity.

The COPS system is implemented as a web application with desktop look-and-feel. With this design the exploration of the structural neighborhood of a given protein structure with the Fold Space Navigator is as easy as browsing through a local file system. Moreover, the Fold Space Navigator contains several tools like qCOPS (quantified COPS) to search the domains in COPS by PDB code or keyword. The returned result list of COPS domains may then be sorted by different criteria like species or PDB header information, and the list can be exported in a variety of text formats. Additionally, the structural similarities of any two domains may be instantly visualized with Jmol (http://www.jmol.org/) and TopMatch (6) by a convenient drag and drop mechanism.

The classification pipeline of COPS is exposed by the iCOPS (instant COPS, (2,7) web application. iCOPS provides a simple upload form for the user to upload a file in PDB format. As soon as the structure has successfully arrived, the classification procedure of iCOPS is triggered. During the classification process the uploaded structure is automatically chopped into domains, the domains are immediately visualized in Jmol, and finally the domains are classified relative to the current COPS hierarchy. The classification of an average protein structure with 300 amino acids takes less than one minute. In the meantime the Fold Space Navigator or any other COPS tool can be used.

The COPS web server is a powerful web application for elaborate explorations in the currently known fold space. It provides a workbench for structural biologists to easily access the structural neighborhood of a given protein structure, and to utilize key technologies like protein structure classification, visualization of pairwise structural similarities, and protein domain decomposition. Availability: http://cops.services.came.sbg.ac.at/.

References

72. COPS-Benchmark: a structure-derived benchmark for homology and fold detection methods.

Markus Gruber, Karl Frank and Manfred Sippl. University of Salzburg, Salzburg, Austria.
We present a normalized benchmark consisting of a set of protein domains of similar structure whose sequence similarity is well below 30%. The benchmark is derived from COPS (Classification Of Protein Structures) and includes a web service that evaluates the performance of any algorithm relative to BLAST, PSI-BLAST and HHsearch.

A frequent task in molecular and structural biology is the detection of homologous proteins and therefore, the design of new algorithms and the fine tuning of existing tools is a major topic in bioinformatics. Here we provide a normalized benchmark to evaluate the ability of any search tool to detect homologies at low sequence identities and to rank the results by means of ROC (receiver operating characteristic) analysis [1,2] and by comparison to established methods.

The COPS-Benchmark presented here is derived from COPS (Classification of Protein Structures) [3,4,5] containing an exhaustive set of structure and sequence relationships of all currently known protein structures. The benchmark consists of 180 groups of structurally related proteins of probable common evolutionary origin. At the same time the pairwise sequence identity among the proteins within a group is less than 30%. Most importantly the benchmark ensures that there are no homologies among the members of different groups. This guarantees that the benchmark does not contain true positives which are counted as false positives in the evaluation. In the COPS-Benchmark each group contains one query and exactly 6 related proteins. In this sense the benchmark is normalized since every group of related proteins contributes exactly the same number of possible true positives. In standard ROC analysis true positives are plotted against the number of false positives. In the analysis represented by the COPS benchmark we have the additional benefit that the result for each query can be analysed independently and that averages can be computed over all groups. This immediately identifies query proteins where particular methods fail. Most importantly we can objectively compare methods, analyse cuts and unions of different methods, and reveal homologies which are exclusively found by one method or another. The COPS-Benchmark service provides the results for several popular homology detection methods, which may be used as a reference (summarized in the Figure). Standard Smith-Waterman algorithms [6] and BLAST [7] perform on the same level, detecting 40% of the homologies. PSI-BLAST [7] performs substantially better with a sensitivity of 70%. The most successful method is HHsearch [8], based on hidden markov models, which detects 90% of all relationships contained in the benchmark. In summary, the COPS-Benchmark provides a convenient and easy to use standard reference for the development of new methods and the fine tuning of existing tools for the identification of homologous proteins.

REFERENCES:

73. An analytical solution for finding voids and bottlenecks within macromolecules

Fredrik Boulund, Viktor Jonsson, Erik Sterner and Graham Kemp, Chalmers University of Technology, Sweden.

We present an implementation of a direct analytical method for finding the largest sphere inscribed by four others. This method has been applied to the identification of voids and bottlenecks in protein channels.

Channels in proteins play important functional roles, enabling the passage of ions and other molecules across a membrane, or providing pathways for the migration of ligands between the protein surface and interior binding sites. Examples of biologically important channels in bacteria include porins in the outer membrane of, for example, Mycobacterium tuberculosis which affect the ability of drugs to pass through the membrane [1], and mechanosensitive channels enabling ion transfer [2]. The increase in the number of channel-containing structures in the Protein Data Bank has been accompanied by the development of computational
methods for finding and analysing channels (e.g. HOLE [3], CAVER [4], MOLE [5], MolAxis [6], HOLLOW [7], CHUNNEL [8]).

We are developing an algorithm to identify channels that makes use of a direct method for finding the largest sphere inscribed by four others, adapted from Langlet [9]. Applying the method to the atoms lining the inside of the protein channel offers a direct analytical measure of the different parts of the channel. The basic output is a set of spheres representing the channel. These spheres, together with the information about which four atoms generate them, can be used as a basis for calculating several interesting channel attributes. Being a general algorithm it can be applied to all sorts of molecules or sets of spheres. A similar approach is used in [10], but in that work an iterative method is used for finding the largest spherical voids [11]. When searching for bottlenecks, defined by triplets of atoms spanning a plane locally perpendicular to the channel, the algorithm gives the sphere inscribed by the three atoms.

REFERENCES:

74. Interface flexibility in hub proteins

Arianna Fornili and Franca Fraternali. King’s College London, London, UK.

The role of flexibility in promoting binding promiscuity of hub proteins is investigated through molecular simulation methods. Preliminary results on the dynamical properties of hubs extracted from the PiSite database are presented and possible correlations with interface diversity are analyzed.

The deep relation between dynamics and function of proteins is now widely acknowledged. Conformational changes are commonly observed in enzymes and they are generally coupled to the interchange between their active and inactive forms, which can be triggered e.g. by substrate binding or phosphorylation. There is a growing evidence that intrinsic mobility is important also in regulating protein-protein interactions [1]. Moreover, conservation of flexibility patterns and in particular of slow, large-amplitude concerted motions has been observed in families of homologous proteins [2].

In particular, dynamical properties seem to play an important role in hub proteins. Intrinsic disorder has been claimed to be one of the peculiar features of hubs, together with low sequence complexity, larger surface and high net charge at the interface [3,4]. However, a complete consensus about the role of disorder in hubs has not yet been achieved [5]. Moreover, the possibility to classify different types of hubs on the basis of the disorder propensity is still under investigation [6].

Here we want to explore the hypothesis that ‘promiscuity’ of binding in hubs can be favored by conformational flexibility, that is by the tendency to sample different regions of the conformational space while keeping a defined structure. Flexibility is a property that can be more easily measured than intrinsic disorder, which imply the absence of a stable structure and which is thus difficult to exactly define and directly detect. A few studies on hub proteins have indeed considered flexibility, but through indirect indexes such as the diversity of the structures found for a given protein in the Protein Data Bank (PDB) [5] or the fraction of residues in loops [3].
We present a preliminary screening of the dynamical properties of a set of hub proteins performed with molecular simulation methods. The analysis of protein motions is coupled with a description of the interfaces and in particular of the different types of binding modes that may occur in hubs. Comparison with the behavior of a reference set of non-hub proteins is also taken into account. Datasets for hubs and non-hubs have been built-up using the PiSite database [7], where proteins from the whole Protein Data Bank are clustered in families with high sequence identity and classified according to the number of binding partners and binding modes. Thresholds on these parameters can be set to distinguish between hub or non-hub proteins [5]. For our datasets we considered only chains without gaps in the backbone, to avoid the necessity of introducing models in the subsequent simulations.

A first fast survey of the dynamical properties of the two datasets has been done using two different approaches, namely tCONCOORD [8] and the Gaussian Network Model [9]. While the first method should allow for a larger exploration of the conformational ensemble around the starting structure, including the prediction of possible conformational transitions, the second one is much faster and it could be applied to much larger datasets if needed. Both approaches provided a measure of the flexibility of the protein through the root mean square fluctuation (r.m.s.f.) with respect to the average structure, together with the atomic covariance matrix from which collective motions could be extracted. Results are being compared to check for consistency or possible complementarity of the two methods.

A structural analysis of the interfaces has been performed with the parameter optimized surface (POPS) method [10,11] by calculating the solvent accessible surface area (SASA) that is buried upon complex formation. The possibility to extract unique interfaces by means of clustering techniques and ASASA profiles is currently being investigated. This would provide a structure-based measure of 'promiscuity' by giving the number of different interfaces that can be exploited by a single hub protein in promoting interactions with partners.

The availability of information on both the protein dynamics and the interface properties will allow to determine possible correlations between flexibility and binding diversity. Analysis is being performed by direct comparison of the r.m.s.f. and binding profiles found for each protein. Possible common patterns either in shape or variance distribution of collective motions of hubs are also being investigated and compared with non-hub proteins.

References

75. QMEANfamily: Integration of evolutionary information in the composite scoring function QMEAN

Pascal Benkert and Torsten Schwede, Swiss Institute of Bioinformatics Basel & Biozentrum, University of Basel, Switzerland.

Scoring functions for the quality estimation of protein structure models are a crucial part of modelling pipelines. Composite scoring functions such as QMEAN have been developed which combine multiple statistical potential terms. Here, we describe how the integration of evolutionary information can be used to improve quality estimation.

Model quality estimation plays a key role in protein structure prediction, since ultimately the accuracy of a model determines its usefulness for specific applications. Typically, a set of alternative models for a given protein is produced in the course of protein structure prediction. By ranking these alternative models based on their estimated quality, one aims to select the best model for the target protein. A variety of scoring functions have been described in
the literature for the derivation of a quality estimate on the basis of single models focusing on different aspects of proteins (e.g. deviation from ideal geometry, evolutionary information, analysis of the alignment, physics-based energy functions and statistical potentials [1]). Recently, we have introduced the composite scoring function QMEAN [2][3], which consists of a linear combination of four statistical potential terms investigating different structural aspects of proteins (i.e. torsion angles, non-bonded interactions on C\_beta or all-atom level and packing) and two terms describing the agreement of predicted and observed secondary structure and solvent accessibility.

Here we present a strategy for combining the composite scoring function with evolutionary information on the target protein family under consideration (QMEANfamily). The basic concept of QMEANfamily is the following: If the model is approximately correct for the target sequence it should also be compatible with proteins belonging to the same family - based on the observation that structure is more conserved than sequence [5]. Vice versa, if the model has an incorrect fold or misaligned regions, then all other sequences of this family should be incompatible with this conformation as well. In order to test this hypothesis, we identified homologous proteins of the same family as the target protein using BLAST [4] and generated “models-based-on-a-model” for these sequences using the initial model of the target protein as structural template. Integrating the quality scores over all models of the protein family should improve the signal in comparison to assessing only a single model.

For each model to be evaluated for the given protein sequence, an ensemble of up to 50 auxiliary models is generated for protein sequences sharing at least 30% sequence identity to the target using the original model as template. The homologous sequences are identified in the NCBI’s non-redundant sequence database clustered at 70% sequence identity. The BLAST alignments are converted into raw models by copying the coordinates of the template (protein backbone and conserved side-chains) without modelling any insertions or non-conserved side-chains. The QMEANfamily score is defined as the average QMEAN score of these auxiliary models covering the protein family.

A prototype version of QMEANfamily has been tested recently as a proof of concept at the quality assessment category of the CASP8 blind test experiment. In terms of correlation between predicted model quality and degree of nativeness (GDT\_TS score) on the model submitted by modelling servers, QMEANfamily performed significantly better than QMEAN (t-test, p=0.0006). For 106 CASP8 targets evaluated by both QMEAN and QMEANfamily, the latter generates better Pearson’s correlations in 98 cases. In comparison with other non-clustering scoring functions operating on single models in CASP8, QMEANfamily performed equally well than two other methods (SAM\_T08\_MQAO and MULTICOM\_CMFR) and was only outperformed by SAM\_T08\_MQAU by K. Karplus and co-workers (t-test on Pearson’s correlation coefficients). The latter method is a combination of the Undertaker scoring function with a scoring function based on distance constraints extracted from various target-template alignments [6].

Averaging model quality estimates over auxiliary models covering the protein family seems to primarily reduce the variance in the correlations of predicted vs. calculated model quality and helps reducing outliers, but only occasionally improves model selection, i.e. the quality of the highest ranking model.

The current implementation of QMEANfamily is only based on structural information from the starting model, and does not take into account information from other models or alternative template structures. Therefore, our approach is appropriate for medium to difficult modelling situations in which only a small number of alternative template structures is available. However, in cases with several alternative template structures, it has been shown recently [6][7][8] that using the degree a model deviates from structural constraints extracted from available templates is a very promising strategy for quality estimation. As this information is already implicitly available in the family multiple sequence alignment, using alternative template structure information would be a natural extension of QMEANfamily in future versions. The QMEAN scoring function is publicly available as part of the QMEAN server [9] under the following URL: http://swissmodel.expasy.org/qmean/

References:

76. ModLink+: improving fold recognition by using protein-protein interactions

The fold of a protein sequence can be predicted using different strategies. Here, we show how protein-protein interaction information can improve the fold recognition capacity of state-of-the-art methodologies. In addition, we show that our method can be applied large-scale (e.g. in yeast).

In the post-genomic era, the comprehension of the structure and function of proteomes is essential for the exploitation of the available biological information. This knowledge is of important relevance for understanding the biological processes and it can help in the future development of therapeutics. In the recent years, the large-scale sequencing methods have provided a vast amount of protein sequences, but even for well-characterized organisms, the structure and function remains unknown for a significant fraction of their proteomes.

There have been some recent structural genomics initiatives and biochemical efforts to characterize the structure and function of proteins. However, the easiest way to gauge the information about the structure and function of a protein is by transferring annotation from characterized proteins of similar sequence. Nevertheless, it is known that the structure of a protein generally provides more information about the function it can be involved in than its sequence alone. Moreover, even for those cases where two proteins have a low sequence similarity, they can share the same structure. Although there are numerous pitfalls on the application of this paradigm, it is agreed that the knowledge of the structure of a protein facilitates the prediction of its function. Therefore, it is important to determine or predict the structures of as many proteins as possible.

The structure of a protein can be frequently divided into one or more domains, which may interact with domains from other proteins. Moreover, homologous proteins tend to interact through similar domains. Thus, it may be possible to predict the fold of a protein from its interacting partners of known structure. We developed a method, named ModLink, which used both sequence similarity and protein-protein interactions to assign a SCOP fold and a family classification to uncharacterized proteins. The rationale behind ModLink is that two proteins are more likely to be homologous if they also have similar interaction partners. Nevertheless, the method could only be applied if the accuracy and number of interactions available for a query protein were high. To overcome this limitation, ModLink increased the available interactions by extrapolation: two proteins were linked by extrapolation if any members from their SCOP families interacted with each other. By design, ModLink was not able to deal successfully with proteins having a large number of different interacting partners, usually referred to as “hubs”. As a result, ModLink did not use such hub proteins for extrapolating links. Therefore, the performance of ModLink would improve if the method could distinguish between hub proteins whose interacting partners have similar sequences and those that do not.

Here, we describe a new version of ModLink, called ModLink+, which includes an improved procedure for extrapolating links. This algorithm iteratively varies the number of interactions required to consider a protein as a hub. The approach, that comprises a “self-adaptive” definition of hub proteins, has increased applicability without affecting its accuracy.

ModLink+ has been compared against the original ModLink, PSI-BLAST, and 2 of the newest and most accurate methods of homology detection: HHSearch and PRC. 3,716 non-redundant protein sequences conformed the test set. These sequences had domains with annotated folds in SCOP (thus, allowing more than one SCOP fold annotation per protein sequence) and known interacting partners in DIP. For this test set, the ratio of success (PPV) on fold assignment increases from 75% for PSI-BLAST, 83% for HHSearch and 81% for PRC to >90% for ModLink+ at the e-value cutoff of 10-3. Under this e-value, ModLink+ can assign a fold to 30-45% of the proteins in the test set (depending of the specifications of the program), while our previous method could cover <25%. Moreover, in order to test the applicability of ModLink+ in a realistic scenario, we assigned putative folds to 6,313 proteins of unknown fold in yeast using all the protein-protein interaction data available. There, ModLink+ assigns the fold for domains in 3,738 proteins, while PSI-BLAST alone covers only 2,122 proteins, HHSearch 2,969 and PRC 2,826 proteins, using a threshold e-value that would represent a PPV >82% for each method in the test set. Thus, if the knowledge on protein-protein interactions was complete, our results in yeast suggest that ModLink+ could be applied to more than 2,000,000 sequences in the UniProt database.

In summary, we have shown that ModLink+ surpasses current methods of remote homology detection in coverage and accuracy. In other words, we improved PSI-BLAST HHSearch and PRC by using protein-protein interactions even at the expense of reducing their original coverage (applicability).

Fig. Total number of proteins with assigned fold versus the threshold on the sequence similarity e-value for ModLink+ (green), ModLink (orange), PSI-BLAST (black), HHSearch (cyan) and PRC (blue). Extrapolation was based on SCOP family codes. Red dots indicate the number of proteins with an assigned fold when applying a threshold on the e-value at which the PPV is >82% (for the test set). The difference between each method of the total number of target proteins with at least one assigned fold is shown in the right margin. The distribution of target proteins with domains with an assigned fold obtained with PSI-BLAST (black), ModLink
Ankyrin repeat proteins (ARPs) consist of repeating motifs named ankyrin repeats, which stack with each other forming a non-globular three-dimensional structure. The tandem arrangement of ankyrin repeats (ARs) in ARPs displays lack of long-range interactions so that these proteins are an ideal paradigm for studying of protein folding and stability. Here we report all-atom explicit water molecular dynamic (MD) simulations (parm99 force field) on a ~3.2 micro-seconds timescale carried out in order to monitor the stability of the ARP p18INK4c (p18) and assess the functional role of its fifty fragments. The fragment simulations show that p18 alpha-helices are stabilized by tertiary interactions, because in the absence of their native context they melt in timescales of several hundred picoseconds. All single p18 ARs are also unstable outside their native context and quickly unfold. On the other hand, pairs of ARs are able to maintain their native structure during the entire simulation time so that we can consider them as the minimal stable units of the p18 protein. Moreover, pairs of internal ARs are less stable than pairs that include a native terminal AR. The MD simulations also provide information about the functional roles of p18 turns and loops; the turns appear to be important for the stability of the p18 protein structure, while the loops both help to stabilize the protein and are involved in recognition.

The support through grants LC512, MSM6198959216, and 203/09/H046 is gratefully acknowledged.
The Protein Model Portal resource of the PSI Structural Genomics Knowledgebase provides a single interface to simultaneously query experimental structures and pre-computed comparative models at various participating sites, and to access interactive services for template selection, target-template alignment, model building, and quality assessment.


The three-dimensional structure of a protein provides important information for understanding its biochemical function and interaction properties in molecular detail. However, the number of known protein sequences is much larger than the number of experimentally solved protein structures. As of April 2009, more than 55,000 experimentally determined protein structures were deposited in the Protein Data Bank (PDB)(1). Yet, this number appears relatively small compared with the almost 8 million protein sequences held in the UniProt knowledgebase(2). Homology (or comparative) modeling methods make use of experimental protein structures to build models for evolutionarily related proteins. This technique predicts the three-dimensional structure of a given protein sequence (target) based primarily on its alignment to one or more proteins of known structure (templates). For every experimentally determined structure, often models for hundreds of proteins can be derived using a variety of established methods for comparative protein structure modeling methods, which dramatically increases the structural coverage of protein sequences. Structural genomics and homology modeling thereby complement each other in the exploration of protein structure space.

Using model information effectively is however not always straightforward due to the fact that model information is heterogeneous, highly context dependent, dynamic, and consists of large volumes of data. Moreover to access all models available for a specific protein, the heterogeneous formats at different databases and provider sites which use various incompatible accession code systems has to be taken into consideration. To facilitate the use of protein structure models we have therefore developed the Protein Model Portal (PMP)(3) which provides a single portal which gives access to all experimental structures and available comparative models of the currently participating sites (CSMP - Center for Structures of Membrane Proteins, JCSG - Joint Center for Structural Genomics, MCGS - Midwest Center for Structural Genomics, NESG - Northeast Center for Structural Genomics, NYSGXRC - NY SGX Research Center for Structural Genomics, JCMM - Joint Center for Molecular Modeling, ModBase – UCSF, SWISS-MODEL Repository, Swiss Institute of Bioinformatics & University of Basel) for a given protein.

A single interface allows querying simultaneously experimental structures and pre-computed models across various sites. Functional annotation for individual target proteins is retrieved in real-time from the annotation providers using web services. The three-dimensional coordinates of the individual models are stored at the different model providers and retrieved by the portal in real-time when required for the visualization of the model. The alignment between the target sequence and the template is inferred dynamically on-the-fly by structural superposition of the final model to the template structure. Models provided by the different participating sites have been generated using various algorithmic approaches with different strengths and weaknesses. Also the quality of individual models highly depends on the evolutionary proximity to the selected structural templates. Users of the portal can access the structural comparison of models from different sources and use tools to identify and visualize structurally conserved and variable regions in an ensemble of models. PMP also provides links to interactive services for template selection, target-template alignment, model building, and quality assessment. The consequent use of portal technologies allows federating a set of heterogeneous resources into a single portal while at the same time ensuring consistency of the exchanged data. The current release of the PMP (May 2009) consists of 7.6 million comparative protein models for 3 million distinct UniProt entries.

The Protein Model Portal has been developed as a resource of the PSI-Nature Structural Genomics Knowledgebase(4). In collaboration between the Protein Structure Initiative (PSI) and Nature Publishing Group (NPG), the knowledgebase covers developments in the fields of structural genomics and structural biology.


80. Extension, distribution and causes of conformational diversity in proteins

Ezequiel Juritz, Sebastian Fernandez Alberti and Gustavo Parisi. Quilmes National University, Argentina.

We studied the extension and distribution of protein conformational diversity as derived from the analysis of CATH database. We found that the distribution of protein conformational diversity is very heterogeneous.
also at the S60 level of homologous superfamilies. We found that this distribution is correlated with functional diversification.

It is well known that the native state of a protein is better described by a set of conformers with about the same energy and in dynamic equilibrium. This conformational diversity is a clue feature in proteins to understand their biological activity and functions and to understand sequence-structure relationship. Since the pioneering experiments of Max Perutz in the early 60s with his studies on the T and R forms of hemoglobin, the study of protein conformations has a central role in several areas of structural biology as functional characterization, drug design and protein promiscuity, development of docking and structural alignment techniques and the understanding of protein evolution.

Here we study the extension, distribution and possible causes of protein conformational diversity in proteins. To this end, we have used proteins with more than one crystallographic structure as derived from the CATH structural database v3.2.0. We collected all the proteins sharing the first 8 codes corresponding to CATH structural classification. We then obtained a total of 49157 structures. In those cases were different chains of a given oligomeric structure were present, a single representative domain were chosen randomly. After this, we obtained 8905 domains were the 45% of them have at least 2 crystallographic structures and up to 34. Using this derived database we estimated the conformational diversity for each domain, using different measures of structural similarity as RMS, TMscore, GDT and Maxsub scores. An all versus all calculation for those structural similarity scores was performed using the structures for each domain in the derived database mentioned above. Then, for each score the average structural dissimilarity was registered and these values were used to estimate conformational diversity.

It could be obvious that repeating the crystallographic studies for a given protein, the same fold and conformation will be obtained. So, to observe different conformations, sample and crystallographic conditions should be changed in order to observe a given diversity in the protein structure. We estimated that the conditions that could conduce to the observation of conformational diversity were presence of ligand/s, variation in the oligomeric structures, changes in temperature or pH, presence of mutations and chemical modification of the protein (for example presence/absence of disulfide bonds). This information was obtained from different databases as PDB and PQS or analyzing proteins structures and complemented with other information such us GO terms and taxonomy for each domain.

A null distribution of domains, those for which the same sample or crystallographic conditions are observed for all the structures retrieved from CATH, was built in the following way. Considering all the structures for a given domain, if any structure contains or show any of the conditions mentioned above (any of the structures present ligands, mutations, oligomeric change and so on) these structure belong to the null distribution. Then, the averaged measure of structural similarity gave us just the conformational diversity of a protein for repeating the same experiment in similar conditions. From the set of 8905 domains, the great majority falls in the null distribution with an average RMS of 0.4. We then considered the domains containing at least one of the factors mentioned above, for example a domain containing at least a structure with pH variation compared with the rest of the structure for the same domain. We argued that if a variation of pH in one of the structure crystallized for a given domain could produce a conformational change; this change should be noticed with the structural comparison parameters mentioned above. To assess this change we compared the distribution for all domains with at least one change (for example sample pH) with the null distribution mentioned above. We established that a domain have conformational diversity if a given parameter of structural similarity measure statistically differ from the null distribution.

In general we found that conformational diversity extents from 1.4 to 20.4 A as the RMS average value obtained among the different structures for a given domain. The extension of conformational diversity does not depend on protein length, on the number of crystallized structures for each protein neither on the presence of mutations. The presence of ligands, variation of pH and variation on the number of disulfide bridges are among the most important factors producing conformational diversity. Using CATH structural classification, each of the three main structural classes (mainly alpha, mainly beta and mixed alpha and beta) seems to have similar content of conformational diversity. However, architectures and topologies for each class showed a clear heterogeneity in conformational diversity extension. At this level we found a strong correlation with functional diversity using GO terms classification of proteins. These results reflect the great diversity found when the homologous superfamily level was evaluated. At this level and in spite of the great structural similarity, the heterogeneity in conformational diversity is observed up the S60 level (homologous families with more than 60% identity).

Our results indicate that the distribution of protein conformational diversity is not uniform in the structural space provided by CATH database. There exists a strong heterogeneity in the extension of protein conformational diversity also within the S60 level. Although several reports indicates that homologous proteins with the same overall fold share their dynamics and structural deformations, here we found that the extension of the conformational diversity reached by a given protein is strongly influenced by functional constraints during evolution.

81. Similarity of Chemical Mechanisms in Functionally Analogous Enzymes


We compared 95 pairs of functionally analogous enzymes (enzymes that catalyze similar chemical transformations but do not share common ancestry) from the MACiE database. We conclude that functional analogues that catalyze similar overall transformations have commonly
converged to use similar catalytic mechanisms, with several pairs sharing identical mechanistic steps.

It has been seen that unrelated genes in different species can evolve to produce enzymes catalyzing similar overall transformations on similar substrates. These enzymes are said to be functionally analogous, and because they are not relevant to genomic enzymology, they have been poorly studied. As a consequence, several questions about how similar transformations can be catalyzed by different structural scaffolds remain. Answering these questions would give insight into the evolutionary constraints imposed by reactions on the enzymes that catalyze them, specifically in their requirement of catalytic residues, enzyme chemistries, binding sites, and ultimately the whole structure of the enzyme.

Here we quantitatively measured the similarity in overall transformations and reaction mechanisms for 95 pairs of functionally analogous enzymes (non-homologous enzymes sharing up to the third level of their EC classification numbers) from the MACiE database [1]. MACiE currently holds 223 reaction mechanisms for enzymes which have both a structure deposited in the PDB and a plausible reaction mechanism published in the literature. To compare the reactions in MACiE, O’Boyle and colleagues developed the first quantitative measure of similarity between reactions based on their explicit mechanisms [2]. Each mechanistic step of every reaction was coded as the set of bond changes (BC method) occurring in the transformation from substrates to products in that step, and also as a 58-character feature fingerprint (FP method). Similarities between all possible combinations of steps of a pair of reactions were calculated using a Tanimoto coefficient (for the BC method) and a normalized Euclidean distance (for the FP method), and stored in a similarity matrix. Then, the reaction sequences were globally aligned using the Needleman-Wunsch algorithm and their mechanistic similarity expressed as a Tanimoto coefficient using the number of steps in each reaction and the Needleman-Wunsch similarity as inputs.

For this work, we continued the development of the BC method to include the search for local alignments of mechanistic steps, to consider reversibility of enzyme reactions, and to allow circular permutation of steps in the reaction sequences. We also measured the similarities of overall reactions, without regard to their mechanistic approach, by computing Tanimoto coefficients using the set of bond changes based on only the transformation from overall substrate(s) to overall product(s). We first assessed whether the sub-subclass level of the EC classification [3], used to define functional similarity in this and other works, was indicative of overall transformation similarity. We then turned to the question: to what extent does similarity of overall transformations imply similarity of reaction mechanisms?

Forty-one pairs of reactions in the dataset had overall transformations whose similarity was significant at the 5% level. For the remaining 54 pairs, which spanned 13 out of the 29 total EC sub-subclasses considered, the overall similarity was not significantly higher to that of random pairs of reactions. The most dramatic example was that of the carbon-oxygen lyases acting on phosphates (EC 4.2.3), for which none of the six pairs of reactions exhibited significant overall reaction similarity. We also looked at the bond changes that were conserved across all members of each sub-subclass and found that conservation of rare bond changes (e.g. cleavage or formation of C-Cl bonds) was a strong indicator of high overall reaction similarity. In general, the results indicated that EC sub-subclasses performed by several different structural scaffolds commonly encompassed a complex mixture of overall transformations, which may be better redefined as different, more stringent, sub-subclasses.

In terms of mechanistic similarity, 24 pairs in the dataset had similarity scores significant at the 5% level. Interestingly, pairs of enzymes from EC sub-subclasses that showed significant overall reaction similarity were likely also to share mechanistic similarity (14 out of 27). Our results also demonstrated that for functionally analogous pairs of enzymes, the overall reaction similarity serves as an upper bound on the mechanistic similarity (see Figure), tying in with the previous observation. We identified 31 identical pairs of mechanistic steps (out of a total 261 aligned steps among the 95 pairs of reactions) spanning 21 different pairs of reactions. Thus, for reactions with similar overall transformations (41), convergence to the same mechanism was common (19), with several of these reactions also sharing identical mechanistic steps (12). We note that the CSA is a superset of MACiE, and thus the work of Gherardini and colleagues on convergence of active sites in the CSA [4] is directly comparable with the results presented here. In short, they reported that only six out of the 95 pairs of enzymes in our dataset presented similarity of active sites, and thus convergence of catalytic mechanism is independent of convergence of active sites.

Our findings have implications in at least three areas of research. First, our analyses suggest that mechanistic similarity should be used, along with homology, to guide functional annotation and prediction, especially since functionally analogous (but non-homologous) enzymes often have key mechanistic steps in common. Secondly, the abundance of cases of functional convergence demonstrates the ability of nature to reach the same solution to a chemical problem from different starting points. This is encouraging for enzyme engineering, and our methods to measure similarity of mechanisms can be used to identify the best templates from which to start engineering a new function. Thirdly, classification of enzymes based on quantitative measures of overall transformations...
is a promising way to better correlate structure and function. This is further supported by the fact that almost half of the pairs that showed overall reaction similarity also shared mechanistic similarity.


82. Quantum States of Biomolecular Breathers. Insights from Semiclassical and Feynman Path Integral Quantization of Protein Solitons.

Alexander Kantardjiev, Sofia University, Bulgaria.

Currently it is considered that nonlinearity effects in protein molecules cause vibrational modes that could last longer and be spatially localized. The essence of our work is to quantize the breathers (solitons) via the application of via bosonization rules, semiclassical, and quantum mechanical Feynman path integral approaches and characterize breathers quantum states.

It is widely accepted that bioenergetic processes in protein molecules – energy supply, transport, storage - rely on the soliton mechanism [1]. Currently it is considered for sure that nonlinearity effects in protein molecules lead to appearance of vibrational modes that could last longer on the typical protein motion time scale and be spatially localized. Temporal and spatial structures of these nonlinear effects are generally referred to as solitons. Another approach is to describe them in terms of topological excitations or discrete breathers. - nondispersive classical excitations.

A fairly simple scheme for their calculation is based on elastic network model (harmonic approximation) that was proved to reflect topological characteristics. An extension is the nonlinear network model which adds anharmonic terms to the molecular Hamiltonian. As suggested by previous simulations we managed (through application of nonlinear network model framework) to show that spatial and temporal localization of energy can occur in hydrogen bond networks in protein molecules. Such discrete breathers appear under certain combination of parameters that enter explicitly in the hamiltonian function. Hereafter the hamiltonian is parametrized via force field constants and spectroscopic data (especially amide bands, but also amino acids side chain signatures). Our subsequent steps that comprise the essence of our work is to quantify the breathers (solitons) via the application of direct (via bosonization rules), semiclassical, and quantum mechanical path integral approaches.

As might be supposed from first principles considerations the quantum analogues of classical breathers (solitons) are bound quantum states characterized by the number of phonons. Following bosonization rules, the lowest states in this treatment (nonlinear terms in the Hamiltonian operator) were shown to be biphonons. Their appearance is a direct consequence of nonlinear contributions. To the best of our knowledge it is the first report for this quantum state in protein hydrogen bonded chains modeled in terms of a Fermi-Pasta-Ulam breather. Higher quantum soliton states are also described but the focus is on the bivibron states which were previously shown to exist in a minimal model of discrete Fermi-Pasta-Ulam breather [2]. The bosonization procedure followed a standard analytical scheme (boson quantization rules): definition of phonon creation and annihilation operators - either in terms of local modes or collective modes. Soliton states with higher quantum number require numerical procedures and some preliminary results are reported.

Semiclassical quantization of protein solitons (breathers) excitations relies on Einstein-Brillouin-Keller (EBK) quantization of the classical action. In our presentation we emphasize the connection of the intuitive and obvious protein structure notions with the formal construction of this powerful quantization scheme. We give explicit view of the 2-KAM (Kolmogorov-Arnold-Moser) tori. In addition, we applied a recently described method [3] for deriving the semiclassical wave function using path integrals. The shape of the wave function can give insights in the quest for experimental work to support theoretical predictions [3]. At last, to strengthen our arguments we just mention that semiclassical methodology is considered to be a reasonable approximation for excited systems (such as protein breathers, solitons).

A fully quantum mechanical approach to protein solitons has two routes: direct diagonalization of the Hamiltonian and Feynmann path integral quantization.

In the protein structure context the first approach boils down to choosing a reasonable cutoff in the phonon Hilbert space as confined by the protein molecule dynamics. Finding an eigenstate of the protein hamiltonian is necessary but not sufficient condition - a check for localization of the eigenstate is needed in order to classify it as a breather. The criteria is the observation that the breather state is comprised of states of an approximately quadratic local mode [3]. Feynman path integral approach allows one to integrate analytically protein harmonic (quadratic) degrees of freedom and thus lowering the dimensionality of the problem. Criteria for localization - decoupling of the oscillating atoms from the protein molecule (upon raising the quartic coupling constant) as if for a quartic local mode.

Throughout we do not touch on the subtle issue of quantum tunneling processes. The argument for their neglect is in the time scale – quantum tunneling is orders slower. One perspective that is discussed is the effect of the protein electrostatics [4,5] - ionization state of the molecule - and therefore hydrogen bond network topology on the soliton formation. We have gathered computational results that suggest the possibility for electrostatic effects in controlling these mechanisms (via hydrogen bond network state) as well as via long range effects of dipole moments [5].

We hope that application of these advanced theoretical tools of quantum mechanics to protein structure and dynamics will help researchers in their effort to observe these effects in real systems (proteins). We express confidence in the adequacy of the inevitable approximations so that insightful hints emerge about the way protein molecules store and transfer energy via vsoliton mechanisms. Hope time has come to fit these theoretical predictions to the results of spectroscopy experiments. On the
fundamental side fields like bioenergetics and quantum consciousness hypothesis (on the basis of quantum coherence in protein molecules) might benefit from these lines of research.

References:
### LIST OF PARTICIPANTS

<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Country</th>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abeln</td>
<td>Sanne</td>
<td>Holland</td>
<td><a href="mailto:s.abeln@amolf.nl">s.abeln@amolf.nl</a></td>
</tr>
<tr>
<td>Abroi</td>
<td>Aare</td>
<td>Estonia</td>
<td><a href="mailto:aabroi@ebc.ee">aabroi@ebc.ee</a></td>
</tr>
<tr>
<td>Akiva</td>
<td>Eyal</td>
<td>Israel</td>
<td><a href="mailto:eyala@ekmd.huji.ac.il">eyala@ekmd.huji.ac.il</a></td>
</tr>
<tr>
<td>Almonacid</td>
<td>Daniel</td>
<td>USA</td>
<td><a href="mailto:daniel.almonacid@ucsf.edu">daniel.almonacid@ucsf.edu</a></td>
</tr>
<tr>
<td>Barth</td>
<td>Patrick</td>
<td>USA</td>
<td><a href="mailto:barthp@u.washington.edu">barthp@u.washington.edu</a></td>
</tr>
<tr>
<td>Ben-Tal</td>
<td>Nir</td>
<td>Israel</td>
<td><a href="mailto:bental@ashtoret.tau.ac.il">bental@ashtoret.tau.ac.il</a></td>
</tr>
<tr>
<td>Berbalk</td>
<td>Christoph</td>
<td>Austria</td>
<td><a href="mailto:Christoph.Berbalk@sbg.ac.at">Christoph.Berbalk@sbg.ac.at</a></td>
</tr>
<tr>
<td>Bernauer</td>
<td>Julie</td>
<td>France</td>
<td><a href="mailto:julie.bernauer@inria.fr">julie.bernauer@inria.fr</a></td>
</tr>
<tr>
<td>Bonati</td>
<td>Laura</td>
<td>Italy</td>
<td><a href="mailto:laura.bonati@unimib.it">laura.bonati@unimib.it</a></td>
</tr>
<tr>
<td>Bordogna</td>
<td>Annalisa</td>
<td>Italy</td>
<td><a href="mailto:a.bordogna3@campus.unimib.it">a.bordogna3@campus.unimib.it</a></td>
</tr>
<tr>
<td>Bordoli</td>
<td>Lorenza</td>
<td>Switzerland</td>
<td><a href="mailto:lorenza.bordoli@unibas.ch">lorenza.bordoli@unibas.ch</a></td>
</tr>
<tr>
<td>Boulund</td>
<td>Fredrik</td>
<td>Sweden</td>
<td><a href="mailto:fredrik.boulund@gmail.com">fredrik.boulund@gmail.com</a></td>
</tr>
<tr>
<td>Bray</td>
<td>Tracey</td>
<td>UK</td>
<td><a href="mailto:t.bray@postgrad.manchester.ac.uk">t.bray@postgrad.manchester.ac.uk</a></td>
</tr>
<tr>
<td>Cazals</td>
<td>Frederic</td>
<td>France</td>
<td><a href="mailto:Frederic.Cazals@sophia.inria.fr">Frederic.Cazals@sophia.inria.fr</a></td>
</tr>
<tr>
<td>Chabbert</td>
<td>Marie</td>
<td>France</td>
<td><a href="mailto:marie.chabbert@univ-angers.fr">marie.chabbert@univ-angers.fr</a></td>
</tr>
<tr>
<td>Davis</td>
<td>Fred</td>
<td>USA</td>
<td><a href="mailto:davis@janelia.hhmi.org">davis@janelia.hhmi.org</a></td>
</tr>
<tr>
<td>De Leeuw</td>
<td>Marina</td>
<td>Israel</td>
<td><a href="mailto:deleeuw@bgu.ac.il">deleeuw@bgu.ac.il</a></td>
</tr>
<tr>
<td>Deane</td>
<td>Charlotte</td>
<td>UK</td>
<td><a href="mailto:deane@stats.ox.ac.uk">deane@stats.ox.ac.uk</a></td>
</tr>
<tr>
<td>Elofsson</td>
<td>Arne</td>
<td>Sweden</td>
<td><a href="mailto:arne@bioinfo.se">arne@bioinfo.se</a></td>
</tr>
<tr>
<td>Fagerberg</td>
<td>Linn</td>
<td>Sweden</td>
<td><a href="mailto:linnfa@kth.se">linnfa@kth.se</a></td>
</tr>
<tr>
<td>Florova</td>
<td>Petra</td>
<td>Czech Republic</td>
<td><a href="mailto:sazk@centrum.cz">sazk@centrum.cz</a></td>
</tr>
<tr>
<td>Fornili</td>
<td>Arianna</td>
<td>UK</td>
<td><a href="mailto:arifor@gmail.com">arifor@gmail.com</a></td>
</tr>
<tr>
<td>Forrest</td>
<td>Lucy</td>
<td>Germany</td>
<td><a href="mailto:lucy.forrest@biophys.mpg.de">lucy.forrest@biophys.mpg.de</a></td>
</tr>
<tr>
<td>Fracalvieri</td>
<td>Domenico</td>
<td>Italy</td>
<td><a href="mailto:d.fracalvieri@campus.unimib.it">d.fracalvieri@campus.unimib.it</a></td>
</tr>
<tr>
<td>Gallin</td>
<td>Warren</td>
<td>Canada</td>
<td><a href="mailto:wgallin@ualberta.ca">wgallin@ualberta.ca</a></td>
</tr>
<tr>
<td>Glanville</td>
<td>Jacob</td>
<td>USA</td>
<td><a href="mailto:jacob.glanville@pfizer.com">jacob.glanville@pfizer.com</a></td>
</tr>
<tr>
<td>Gofman</td>
<td>Yana</td>
<td>Israel</td>
<td><a href="mailto:gofmany@post.tau.ac.il">gofmany@post.tau.ac.il</a></td>
</tr>
<tr>
<td>Gruber</td>
<td>Markus</td>
<td>Austria</td>
<td><a href="mailto:mgruber@came.sbg.ac.at">mgruber@came.sbg.ac.at</a></td>
</tr>
<tr>
<td>Gursoy</td>
<td>Attila</td>
<td>Turkey</td>
<td><a href="mailto:agursoy@ku.edu.tr">agursoy@ku.edu.tr</a></td>
</tr>
<tr>
<td>Hamelryck</td>
<td>Thomas</td>
<td>Denmark</td>
<td><a href="mailto:thomas.hamelryck@gmail.com">thomas.hamelryck@gmail.com</a></td>
</tr>
<tr>
<td>Hegyi</td>
<td>Hedi</td>
<td>Hungary</td>
<td><a href="mailto:hegyihedi@gmail.com">hegyihedi@gmail.com</a></td>
</tr>
<tr>
<td>Heifets</td>
<td>Abraham</td>
<td>Canada</td>
<td><a href="mailto:aheifets@cs.toronto.edu">aheifets@cs.toronto.edu</a></td>
</tr>
<tr>
<td>Heuser</td>
<td>Philipp</td>
<td>Germany</td>
<td><a href="mailto:philipp.heuser@embl-hamburg.de">philipp.heuser@embl-hamburg.de</a></td>
</tr>
<tr>
<td>Illergard</td>
<td>Kristoffer</td>
<td>Sweden</td>
<td><a href="mailto:kriil@sbc.su.se">kriil@sbc.su.se</a></td>
</tr>
<tr>
<td>Name</td>
<td>Surname</td>
<td>Country</td>
<td>Email</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------</td>
<td>---------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Jegerschold</td>
<td>Caroline</td>
<td>Sweden</td>
<td><a href="mailto:caroline.jegerschold@ki.se">caroline.jegerschold@ki.se</a></td>
</tr>
<tr>
<td>Jonsson</td>
<td>Viktor</td>
<td>Sweden</td>
<td><a href="mailto:v.a.jonsson@gmail.com">v.a.jonsson@gmail.com</a></td>
</tr>
<tr>
<td>Kamichetty</td>
<td>Hetunandan</td>
<td>USA</td>
<td><a href="mailto:hetu@cs.cmu.edu">hetu@cs.cmu.edu</a></td>
</tr>
<tr>
<td>Keating</td>
<td>Kevin</td>
<td>USA</td>
<td><a href="mailto:Kevin.Keating@yale.edu">Kevin.Keating@yale.edu</a></td>
</tr>
<tr>
<td>Kelm</td>
<td>Sebastian</td>
<td>UK</td>
<td><a href="mailto:kelm@stats.ox.ac.uk">kelm@stats.ox.ac.uk</a></td>
</tr>
<tr>
<td>Kemp</td>
<td>Graham</td>
<td>Sweden</td>
<td><a href="mailto:kemp@chalmers.se">kemp@chalmers.se</a></td>
</tr>
<tr>
<td>Keskin</td>
<td>Ozlem</td>
<td>Turkey</td>
<td><a href="mailto:okeskin@ku.edu.tr">okeskin@ku.edu.tr</a></td>
</tr>
<tr>
<td>Khatib</td>
<td>Firas</td>
<td>USA</td>
<td><a href="mailto:firas@u.washington.edu">firas@u.washington.edu</a></td>
</tr>
<tr>
<td>Kiefer</td>
<td>Florian</td>
<td>Switzerland</td>
<td><a href="mailto:florian.kiefer@stud.unibas.ch">florian.kiefer@stud.unibas.ch</a></td>
</tr>
<tr>
<td>Knapp</td>
<td>Bernhard</td>
<td>Austria</td>
<td><a href="mailto:bernhard.knapp@meduniwien.ac.at">bernhard.knapp@meduniwien.ac.at</a></td>
</tr>
<tr>
<td>Kokoszynska</td>
<td>Katarzyna</td>
<td>Poland</td>
<td><a href="mailto:kasia@integromics.pl">kasia@integromics.pl</a></td>
</tr>
<tr>
<td>Kosinski</td>
<td>Jan</td>
<td>Poland</td>
<td><a href="mailto:kosa@genesilico.pl">kosa@genesilico.pl</a></td>
</tr>
<tr>
<td>Kundrotas</td>
<td>Petras</td>
<td>USA</td>
<td><a href="mailto:pkundro@ku.edu">pkundro@ku.edu</a></td>
</tr>
<tr>
<td>Labudde</td>
<td>Dirk</td>
<td>Germany</td>
<td><a href="mailto:dirk.labudde@biotec.tu-dresden.de">dirk.labudde@biotec.tu-dresden.de</a></td>
</tr>
<tr>
<td>Lackner</td>
<td>Peter</td>
<td>Austria</td>
<td><a href="mailto:Peter.Lackner@sbg.ac.at">Peter.Lackner@sbg.ac.at</a></td>
</tr>
<tr>
<td>Laerdahl</td>
<td>Jon K</td>
<td>Norway</td>
<td><a href="mailto:jonkl@medisin.uio.no">jonkl@medisin.uio.no</a></td>
</tr>
<tr>
<td>Lance</td>
<td>Braddon</td>
<td>Australia</td>
<td><a href="mailto:blance@efs.mq.edu.au">blance@efs.mq.edu.au</a></td>
</tr>
<tr>
<td>Lasker</td>
<td>Keren</td>
<td>Israel</td>
<td><a href="mailto:kerenl@post.tau.ac.il">kerenl@post.tau.ac.il</a></td>
</tr>
<tr>
<td>Lathrop</td>
<td>Richard</td>
<td>USA</td>
<td><a href="mailto:rickl@uci.edu">rickl@uci.edu</a></td>
</tr>
<tr>
<td>Levitt</td>
<td>Michael</td>
<td>USA</td>
<td><a href="mailto:michael.levitt@stanford.edu">michael.levitt@stanford.edu</a></td>
</tr>
<tr>
<td>Lindahl</td>
<td>Erik</td>
<td>Sweden</td>
<td><a href="mailto:lindahl@cb.rusu.se">lindahl@cb.rusu.se</a></td>
</tr>
<tr>
<td>Lopez</td>
<td>Daniel</td>
<td>Spain</td>
<td><a href="mailto:dlopez@cnb.csic.es">dlopez@cnb.csic.es</a></td>
</tr>
<tr>
<td>Madera</td>
<td>Martin</td>
<td>UK</td>
<td><a href="mailto:martin.madera@gmail.com">martin.madera@gmail.com</a></td>
</tr>
<tr>
<td>Malisi</td>
<td>Christoph</td>
<td>Germany</td>
<td><a href="mailto:christoph.malisi@tuebingen.mpg.de">christoph.malisi@tuebingen.mpg.de</a></td>
</tr>
<tr>
<td>Mandel-Gutfreund</td>
<td>Yael</td>
<td>Israel</td>
<td><a href="mailto:yaelmg@tx.technion.ac.il">yaelmg@tx.technion.ac.il</a></td>
</tr>
<tr>
<td>Mares-Samano</td>
<td>Sergio</td>
<td>UK</td>
<td><a href="mailto:aveazulms@yahoo.com">aveazulms@yahoo.com</a></td>
</tr>
<tr>
<td>Margraf</td>
<td>Thomas</td>
<td>Germany</td>
<td><a href="mailto:margraf@zbh.uni-hamburg.de">margraf@zbh.uni-hamburg.de</a></td>
</tr>
<tr>
<td>Mashiach</td>
<td>Efrat</td>
<td>Israel</td>
<td><a href="mailto:efratmas@post.tau.ac.il">efratmas@post.tau.ac.il</a></td>
</tr>
<tr>
<td>Mavridis</td>
<td>Lazaros</td>
<td>France</td>
<td><a href="mailto:lazaro.mavridis@loria.fr">lazaro.mavridis@loria.fr</a></td>
</tr>
<tr>
<td>Mehio</td>
<td>Wissam</td>
<td>UK</td>
<td><a href="mailto:w.mehio@ed.ac.uk">w.mehio@ed.ac.uk</a></td>
</tr>
<tr>
<td>Mendez</td>
<td>Ricardo</td>
<td>USA</td>
<td><a href="mailto:ricardo.gonzales7@upr.edu">ricardo.gonzales7@upr.edu</a></td>
</tr>
<tr>
<td>Meirovitch</td>
<td>Hagai</td>
<td>USA</td>
<td><a href="mailto:hagaim@pitt.edu">hagaim@pitt.edu</a></td>
</tr>
<tr>
<td>Melo</td>
<td>Francisco</td>
<td>Chile</td>
<td><a href="mailto:fmelo@bio.puc.cl">fmelo@bio.puc.cl</a></td>
</tr>
<tr>
<td>Mitchell</td>
<td>Julie</td>
<td>USA</td>
<td><a href="mailto:jcmitchell@wisc.edu">jcmitchell@wisc.edu</a></td>
</tr>
<tr>
<td>Moulte</td>
<td>John</td>
<td>USA</td>
<td><a href="mailto:jmoulte@tunc.org">jmoulte@tunc.org</a></td>
</tr>
<tr>
<td>Mueller</td>
<td>Klaus</td>
<td>USA</td>
<td><a href="mailto:klaus.mueller@roche.com">klaus.mueller@roche.com</a></td>
</tr>
<tr>
<td>Najmanovich</td>
<td>Rafael</td>
<td>Canada</td>
<td><a href="mailto:rafael.najmanovich@usherbrooke.ca">rafael.najmanovich@usherbrooke.ca</a></td>
</tr>
<tr>
<td>Nordling</td>
<td>Erik</td>
<td>Sweden</td>
<td><a href="mailto:erik.nordling@biovitrum.com">erik.nordling@biovitrum.com</a></td>
</tr>
<tr>
<td>O'Donoghue</td>
<td>Sean</td>
<td>Germany</td>
<td><a href="mailto:sean@mandala.cc">sean@mandala.cc</a></td>
</tr>
<tr>
<td>Name</td>
<td>Surname</td>
<td>Country</td>
<td>Email</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
<td>--------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Ozkirimli</td>
<td>Elif</td>
<td>Turkey</td>
<td><a href="mailto:elif.ozkirimli@boun.edu.tr">elif.ozkirimli@boun.edu.tr</a></td>
</tr>
<tr>
<td>Pandini</td>
<td>Alessandro</td>
<td>UK</td>
<td><a href="mailto:alessandro.pandini@gmail.com">alessandro.pandini@gmail.com</a></td>
</tr>
<tr>
<td>Parisi</td>
<td>Gustavo</td>
<td>Argentina</td>
<td><a href="mailto:gustavo@unq.edu.ar">gustavo@unq.edu.ar</a></td>
</tr>
<tr>
<td>Ponomarenko</td>
<td>Julia</td>
<td>USA</td>
<td><a href="mailto:jpon@sdsce.edu">jpon@sdsce.edu</a></td>
</tr>
<tr>
<td>Qiu</td>
<td>Yang</td>
<td>USA</td>
<td><a href="mailto:yang.x.qiu@gsk.com">yang.x.qiu@gsk.com</a></td>
</tr>
<tr>
<td>Reynolds</td>
<td>Christopher</td>
<td>UK</td>
<td><a href="mailto:christopher.reynolds08@imperial.ac.uk">christopher.reynolds08@imperial.ac.uk</a></td>
</tr>
<tr>
<td>Sauer</td>
<td>Uwe</td>
<td>Sweden</td>
<td><a href="mailto:uwe.sauer@chem.umu.se">uwe.sauer@chem.umu.se</a></td>
</tr>
<tr>
<td>Savol</td>
<td>Andrej</td>
<td>USA</td>
<td><a href="mailto:ajs171@pitt.edu">ajs171@pitt.edu</a></td>
</tr>
<tr>
<td>Schroeder</td>
<td>Michael</td>
<td>Germany</td>
<td><a href="mailto:ms@biotec.tu-dresden.de">ms@biotec.tu-dresden.de</a></td>
</tr>
<tr>
<td>Schueler-Furman</td>
<td>Ora</td>
<td>Israel</td>
<td><a href="mailto:oraf@ekmd.huji.ac.il">oraf@ekmd.huji.ac.il</a></td>
</tr>
<tr>
<td>Schwede</td>
<td>Torsten</td>
<td>Switzerland</td>
<td><a href="mailto:torsten.schwede@unibas.ch">torsten.schwede@unibas.ch</a></td>
</tr>
<tr>
<td>Sen</td>
<td>Taner</td>
<td>USA</td>
<td><a href="mailto:taner@iastate.edu">taner@iastate.edu</a></td>
</tr>
<tr>
<td>Shazman</td>
<td>Shula</td>
<td>Israel</td>
<td><a href="mailto:shulas@technix.technion.ac.il">shulas@technix.technion.ac.il</a></td>
</tr>
<tr>
<td>Schcushan</td>
<td>Maya</td>
<td>Israel</td>
<td><a href="mailto:minimayas@gmail.com">minimayas@gmail.com</a></td>
</tr>
<tr>
<td>Simon</td>
<td>Istvan</td>
<td>Hungary</td>
<td><a href="mailto:simon@enzim.hu">simon@enzim.hu</a></td>
</tr>
<tr>
<td>Sklenovsky</td>
<td>Petr</td>
<td>Czech Republic</td>
<td><a href="mailto:petr.sklenovsky@seznam.cz">petr.sklenovsky@seznam.cz</a></td>
</tr>
<tr>
<td>Soga</td>
<td>Shinji</td>
<td>Japan</td>
<td><a href="mailto:shinji.soga@jp.astellas.com">shinji.soga@jp.astellas.com</a></td>
</tr>
<tr>
<td>Sontheimer</td>
<td>Jana</td>
<td>Germany</td>
<td><a href="mailto:jana.sontheimer@biotec.tu-dresden.de">jana.sontheimer@biotec.tu-dresden.de</a></td>
</tr>
<tr>
<td>Srivastava</td>
<td>Sheenal</td>
<td>Australia</td>
<td><a href="mailto:Sheenal.Srivastava@efs.mq.edu.au">Sheenal.Srivastava@efs.mq.edu.au</a></td>
</tr>
<tr>
<td>Sternberg</td>
<td>Michael</td>
<td>UK</td>
<td><a href="mailto:m.sternberg@imperial.ac.uk">m.sternberg@imperial.ac.uk</a></td>
</tr>
<tr>
<td>Sterner</td>
<td>Erik</td>
<td>Sweden</td>
<td><a href="mailto:erik.sterner@gmail.com">erik.sterner@gmail.com</a></td>
</tr>
<tr>
<td>Suhrer Jr.</td>
<td>Stefan</td>
<td>Austria</td>
<td><a href="mailto:stefan.suhrer@came.sbg.ac.at">stefan.suhrer@came.sbg.ac.at</a></td>
</tr>
<tr>
<td>Tai</td>
<td>Chin-Hsien (Emily)</td>
<td>USA</td>
<td><a href="mailto:taic@mail.nih.gov">taic@mail.nih.gov</a></td>
</tr>
<tr>
<td>Thiltgen</td>
<td>Grant</td>
<td>USA</td>
<td><a href="mailto:thiltgen@soe.ucsc.edu">thiltgen@soe.ucsc.edu</a></td>
</tr>
<tr>
<td>Togawa</td>
<td>Roberto</td>
<td>Brasil</td>
<td>togawa@cenargen embrapa.br</td>
</tr>
<tr>
<td>Tomczak</td>
<td>Aurelie</td>
<td>Germany</td>
<td><a href="mailto:aurelie.tomczak@biotec.tu-dresden.de">aurelie.tomczak@biotec.tu-dresden.de</a></td>
</tr>
<tr>
<td>Tosatto</td>
<td>Silvio</td>
<td>Italy</td>
<td><a href="mailto:silvio@crib.uni.pd.it">silvio@crib.uni.pd.it</a></td>
</tr>
<tr>
<td>Valentini</td>
<td>Maria</td>
<td>Italy</td>
<td><a href="mailto:maria@crs4.it">maria@crs4.it</a></td>
</tr>
<tr>
<td>Wallach</td>
<td>Izhar</td>
<td>Canada</td>
<td><a href="mailto:izharw@cs.toronto.edu">izharw@cs.toronto.edu</a></td>
</tr>
<tr>
<td>Wass</td>
<td>Mark</td>
<td>UK</td>
<td><a href="mailto:mark.wass04@imperial.ac.uk">mark.wass04@imperial.ac.uk</a></td>
</tr>
<tr>
<td>Wiegels</td>
<td>Tim</td>
<td>Germany</td>
<td><a href="mailto:wiegels@embl-hamburg.de">wiegels@embl-hamburg.de</a></td>
</tr>
<tr>
<td>Wolfson</td>
<td>Haim</td>
<td>Israel</td>
<td><a href="mailto:wolfson@post.tau.ac.il">wolfson@post.tau.ac.il</a></td>
</tr>
<tr>
<td>Wood</td>
<td>Graham</td>
<td>Australia</td>
<td><a href="mailto:gwood@efs.mq.edu.au">gwood@efs.mq.edu.au</a></td>
</tr>
<tr>
<td>Wouters</td>
<td>Merridee</td>
<td>Australia</td>
<td><a href="mailto:m.wouters@victorchang.unsw.edu.au">m.wouters@victorchang.unsw.edu.au</a></td>
</tr>
<tr>
<td>Yeheskel</td>
<td>Adva</td>
<td>Israel</td>
<td><a href="mailto:szueadv@yahoo.tau.ac.il">szueadv@yahoo.tau.ac.il</a></td>
</tr>
<tr>
<td>Yeturu</td>
<td>Kalidas</td>
<td>India</td>
<td><a href="mailto:ykalidas@gmail.com">ykalidas@gmail.com</a></td>
</tr>
<tr>
<td>Zaslavskiy</td>
<td>Mikhail</td>
<td>France</td>
<td><a href="mailto:mikhail.zaslavskiy@gmail.com">mikhail.zaslavskiy@gmail.com</a></td>
</tr>
<tr>
<td>Zoppe</td>
<td>Monica</td>
<td>Italy</td>
<td><a href="mailto:mzoppe@ifc.cn.it">mzoppe@ifc.cn.it</a></td>
</tr>
<tr>
<td>Zotenko</td>
<td>Elena</td>
<td>Germany</td>
<td><a href="mailto:ezotenko@mpi-inf.mpg.de">ezotenko@mpi-inf.mpg.de</a></td>
</tr>
</tbody>
</table>