# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>2</td>
</tr>
<tr>
<td>Keynote abstracts</td>
<td>4</td>
</tr>
<tr>
<td>List of abstracts</td>
<td>6</td>
</tr>
<tr>
<td>Oral presentations</td>
<td>15</td>
</tr>
<tr>
<td>Poster presentations</td>
<td>70</td>
</tr>
</tbody>
</table>
# PROGRAM

**3Dsig: Day 1** (Saturday, July 22th)

<table>
<thead>
<tr>
<th>Time</th>
<th>ID</th>
<th>Title</th>
<th>Presenting Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00</td>
<td></td>
<td>Opening remarks, Rafael Najmanovich – University of Montreal</td>
<td></td>
</tr>
<tr>
<td>10:20</td>
<td>76</td>
<td>Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability</td>
<td>Abdullah Kahraman</td>
</tr>
<tr>
<td>10:40</td>
<td>25</td>
<td>How nature builds electrostatic interactions in natural enzymes: What can we learn for enzyme design?</td>
<td>Mary Jo Ondrechen</td>
</tr>
<tr>
<td>11:00</td>
<td>28</td>
<td>Computational design of a symmetrical beta-trefoil lectin with cancer cell binding activity</td>
<td>Kam Y. J. Zhang</td>
</tr>
<tr>
<td>11:20</td>
<td>152</td>
<td>CATS (Coordinates of Atoms by Taylor Series): Protein design with backbone flexibility in all locally feasible directions</td>
<td>Mark Hallen and Bruce Donald</td>
</tr>
<tr>
<td>11:40</td>
<td>K1</td>
<td>Computational Protein Design: Judge the protein by the cover, story and taste</td>
<td>Ilan Samish, Amai Proteins</td>
</tr>
<tr>
<td>12:30</td>
<td></td>
<td><strong>LUNCH: Exhibition / Poster area</strong></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>174</td>
<td>Large-scale structure prediction enabled by reliable model quality assessment and improved contact predictions for small families.</td>
<td>Mirco Michel, David Menendez Hurtado and Arne Elofsson</td>
</tr>
<tr>
<td>14:20</td>
<td>77</td>
<td>Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins</td>
<td>Wim Vranken</td>
</tr>
<tr>
<td>14:40</td>
<td>80</td>
<td>Density-based clustering in structural bioinformatics: application to beta turns and antibody CDRs</td>
<td>Roland Dunbrack</td>
</tr>
<tr>
<td>15:00</td>
<td>54</td>
<td>Proteins from Peptides</td>
<td>Andrei N. Lupas</td>
</tr>
<tr>
<td>15:20</td>
<td>66</td>
<td>Improving fragment assembly protein structure prediction</td>
<td>Charlotte Deane</td>
</tr>
<tr>
<td>15:40</td>
<td>73</td>
<td>MESHII-score a method for estimation of protein model accuracy</td>
<td>Chen Keasar</td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td><strong>Coffee break with exhibitors</strong></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>K2</td>
<td>Protein bioinformatics of low resolution structural data</td>
<td>Daisuke Kihara, Purdue University</td>
</tr>
<tr>
<td>17:20</td>
<td>11</td>
<td>EncoMPASS: An Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry</td>
<td>Edoardo Sarti</td>
</tr>
<tr>
<td>17:40</td>
<td>43</td>
<td>Folding membrane proteins by deep transfer learning</td>
<td>Jinbo Xu</td>
</tr>
<tr>
<td>18:00</td>
<td></td>
<td><strong>Poster presentations</strong></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>ID</td>
<td>Title</td>
<td>Presenting Author</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>10:00</td>
<td>136</td>
<td>Deep learning based subdivision approach for large scale macromolecules structure recovery from electron cryo tomograms</td>
<td>Min Xu, Xiaqi Chai, Hariank Muthakana, Xiaodan Liang, Ge Yang, Tzviya Zeev-Ben-Mordehai and Eric Xing</td>
</tr>
<tr>
<td>10:20</td>
<td>10</td>
<td>Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone</td>
<td>Juan Rodriguez-Rivas</td>
</tr>
<tr>
<td>10:40</td>
<td>68</td>
<td>Automated evaluation of quaternary structures from protein crystal structures</td>
<td>Jose Duarte</td>
</tr>
<tr>
<td>11:00</td>
<td>69</td>
<td>Deep Learning in text mining for protein docking using full-text articles</td>
<td>Varsha D. Badal</td>
</tr>
<tr>
<td>11:20</td>
<td>177</td>
<td>SnapDock - Template Based Docking by Geometric Hashing</td>
<td>Michael Estrin and Haim J. Wolfson</td>
</tr>
<tr>
<td>11:40</td>
<td>K3</td>
<td>Improving cancer chemotherapy with structure-based drug repositioning</td>
<td>Michael Schroeder, TU Dresden</td>
</tr>
<tr>
<td>12:30</td>
<td></td>
<td>LUNCH: Exhibition / Poster area</td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>42</td>
<td>PRODIGY: a structure-based method for the prediction of binding affinity in biomolecular complexes</td>
<td>Anna Vangone</td>
</tr>
<tr>
<td>14:20</td>
<td>57</td>
<td>Identifying Multiple Active Conformations of G Protein-Coupled Receptors Using Focused Conformational Sampling</td>
<td>Ravinder Abrol</td>
</tr>
<tr>
<td>14:40</td>
<td>71</td>
<td>The Impact of Conformational Entropy on the Accuracy of the Molecular Docking Software FlexAID in Binding Mode Prediction</td>
<td>Louis-Philippe Morency</td>
</tr>
<tr>
<td>15:00</td>
<td>83</td>
<td>Interactome based drug design based on disease-disease relationships</td>
<td>Gaurav Chopra</td>
</tr>
<tr>
<td>15:20</td>
<td>D1</td>
<td>Will data sciences approaches impact our science? (Discussion animated by Phil Bourne, U. Virginia)</td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td>Coffee break with exhibitors</td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>18</td>
<td>Three-dimensional organisation of human genome</td>
<td>Dariusz Plewczynski</td>
</tr>
<tr>
<td>16:50</td>
<td>23</td>
<td>From Mutations to Mechanisms and Dysfunction via Computation and Mining of Protein Energy Landscapes</td>
<td>Amarda Shehu</td>
</tr>
<tr>
<td>17:10</td>
<td>59</td>
<td>What can human variation tell us about proteins?</td>
<td>Stuart A. MacGowan</td>
</tr>
<tr>
<td>17:30</td>
<td></td>
<td>Closing remarks, Rafael Najmanovich – University of Montreal</td>
<td></td>
</tr>
<tr>
<td>18:00</td>
<td></td>
<td>Poster presentations</td>
<td></td>
</tr>
</tbody>
</table>
KEYNOTE ABSTRACTS

K1. Computational Protein Design: Judge the protein by the cover, story and taste

Ilan Samish
Amai Proteins. Israel.

Computational protein design (CPD), a yet evolving field, includes computer-aided engineering of amino-acid sequences for the partial modification or full de novo design of proteins of interest. The designs are defined by a requested structure, function, or working environment. Next, the protein is designed to the requested target in an iterative and often hierarchical approach. Not less important is the negative design aspect in which the CPD is directed to avoid unwanted designs. Integrating these aspects in a case-study approach is aimed to present the plethora of approaches within the CPD field as well as direct researchers to future challenges. These include advancing the field for the benefit of understanding protein structure and function and the relationships between them as well as applying such know-how for the benefit of mankind as part of the biotechnological industry. Applied aspects range from new biological drugs, via healthier and tastier food products to nanotechnology and environmentally-friendly enzymes replacing toxic chemical reactions utilized in the current industry.


K2. Protein bioinformatics of low resolution structural data

Daisuke Kihara
Purdue University, United States

For many years protein structure bioinformatics has been using protein structures in PDB to elucidate structures and functions of proteins and developing computational methods for the analyses. Although PDB remains as the main source of biomolecular structure data, the game-changing technology development occurred for electron microscopy (EM) in recent years enabled solving macromolecular structures at near atomic resolution using EM. An increasing number of structures determined by EM at various resolutions, from about 1.5 to over 20 Angstroms, are accumulated in EMDB. EM data pose new challenges and exciting
opportunities to the protein bioinformatics community. We will start by overviewing computational methods needed for interpreting EM data of macromolecular structures. Then we will discuss our recent analysis of protein structures determined by EM, and further present methods we developed, including EM-SURFER, which is a server for rapid EMDB search, and structure modeling methods for EM maps.

K3. Improving cancer chemotherapy with structure-based drug repositioning

Michael Schroeder

TU Dresden, Germany

Drug resistance is an important open problem in cancer treatment. In recent years, the heat shock protein HSP27 (HSPB1) was identified as a key player driving resistance development. HSP27 is overexpressed in many cancer types and influences cellular processes such as apoptosis, DNA repair, recombination, and formation of metastases. As a result cancer cells are able to suppress apoptosis and develop resistance to cytostatic drugs.

To identify HSP27 inhibitors we follow a novel structure-based drug repositioning approach. We exploit a similarity between a predicted HSP27 binding site to a viral thymidine kinase to generate lead inhibitors for HSP27. We characterise binding of a known inhibitor with interactions patterns of our tool Plip and exploit this knowledge to assess better binders.

Six of these leads were verified experimentally. They bind HSP27 and down-regulate its chaperone activity. Most importantly, all six compounds inhibit development of drug resistance in cellular assays. One of the leads – chlorpromazine – is an antipsychotic, which has a positive effect on survival time in human breast cancer. The identified compounds will now undergo preclinical studies.

D1. How Will Data Science Influence What We Do?

Philip E. Bourne

University of Virginia, United States

Abstract: Data science is becoming increasingly influential in many industries. Since the research that is being undertaken by those who attend 3Dsig has always been data driven, is there anything new to us emerging from data science and so-called “Big Data?” If the answer is yes, what is new, how is it being applied and what is next? This will be an audience discussion around these questions and other questions that will undoubtedly arise.
LIST OF ABSTRACTS$^{1,2,3}$

1 - Inhibitors binding to ATPase domain of Glucose-regulated Protein 78 affects substrate-binding domain movement: Molecular dynamics studies
Dr Seema Mishra

2 - Uncovering complex structural relationships of proteins
Aleksandar Poleksic

3 - Network approach integrates 3D structural and sequence data to improve protein structural comparison
Khalique Newaz

4 - Drug search for leishmaniasis: a structure-based drug discovery approach for detecting anti-Leishmania hits
Rodrigo Ochoa

5 - Deep learning based subdivision approach for large scale macromolecules structure recovery from electron cryo tomograms
Min Xu

6 - Use of cross-docking simulations for identification of protein-protein interactions sites: the case of proteins with multiple binding sites
Nathalie Lagarde

7 - Using Ancestral Sequence Reconstruction to Characterize an Allosteric Bi-Enzyme Complex
Kristina Heyn

8 - Structure-based drug repositioning identifies novel Hsp27 inhibitors, which efficiently suppress drug resistance development in cancer cells
Michael Schroeder

9 - Zipping and assembly with limited sets of constraints
Maryana Wånggren

---

$^1$ Please note that ID numbers do not correspond to ISMB ID numbers.
$^2$ Full list of authors and affiliations appears in the individual abstracts.
$^3$ Only abstracts submitted via four waves appear in this booklet, those submitted exclusively through ISMB can be found at: http://bit.ly/2sIUzNW.
10 - Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone
Juan Rodriguez-Rivas

11 - EncoMPASS: an Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry
Edoardo Sarti

12 - Recurrent neural network models to quantitatively predict RNA-RNA interactions
Michelle Wu

13 - SPRINT: Ultrafast protein-protein interaction prediction of the entire human interactome
Yiwei Li

14 - Seeing the Trees through the Forest: Sequence-based Homo- and Heteromeric Protein-protein Interaction sites prediction using Random Forest
K. Anton Feenstra

15 - A computational design strategy for discovering the cellular targets of histone lysine methyltransferases
Diego Alonso-Martinez

16 - Engineering Improvement of a Potent Human-Derived Monoclonal Antibody Against Respiratory Syncytial Virus Using Structure-Based Computer Modeling
Sean Le

17 - An efficient algorithm for improving structure-based prediction of transcription factor binding sites
Jun-tao Guo

18 - Three-dimensional organisation of human genome
Dariusz Plewczynski

19 - Structural Determination of Chitinase obtained from Bacteria from the African Catfish (Clarias gariepinus)
Ajayi A.A.

20 - Drug-target interactoin similarity for drug-target interaction prediction and drug repositioning
Daniele Parisi
21 - Frequent Subgraph Mining for Biologically Meaningful Structural Motifs
Sebastian Keller

22 - DeepSF: deep convolutional neural network for mapping protein sequences to folds
Jianlin Cheng

23 - From Mutations to Mechanisms and Dysfunction via Computation and Mining of Protein Energy Landscapes
Amarda Shehu

24 - Understanding Protein Interactions at a Molecular Level with Web Tools Using Inter-Residue Contacts and Intermolecular Contact Maps
Romina Oliva

25 - How nature builds electrostatic interactions in natural enzymes: What can we learn for enzyme design?
Mary Jo Ondrechen

26 - A covalent docking approach to simulate the interactions of a novel cephalosporin derivative with beta-lactamases from different sources
Anna Marabotti

27 - CATH-based protein structure and function analyses to understand the implication of alternative splicing
Christine Orengo

28 - Computational design of a symmetrical beta-trefoil lectin with cancer cell binding activity
Kam Y. J. Zhang

29 - Understanding the Molecular Consequences of Genomic Variation Associated with Drug Resistance in Mycobacterium tuberculosis
Nicholas Furnham

30 - RIP-MD: Generation and Analysis of Residue Interaction Networks in Molecular Dynamics of Proteins
Sebastian Contreras-Riquelme

31 - Combining Sequence and Structural features lead to accurate interpretation of genetic variation: large scale modeling and classification of de novo human mutations with VIPUR
Richard Bonneau
32 - Development of new Rosetta energy functions and rotamer libraries for modeling and design of hybrid systems containing proteins, non-canonical amino acids and peptidomimetic backbones
Richard Bonneau

33 - Unifying genomics and molecular biology data
J. Segura*

34 - On the turning away
Alexandre G. de Brevern

35 - A complete Web resource for Galactosemia-related proteins
Anna Marabotti

36 - DisProt 7.0: a major update of the database of disordered proteins
Silvio Tosatto

37 - 40-fold increase in coverage of structure-based annotations for UniProt entries via the SIFTS resource
Jose M. Dana

38 - Characterization of the GPR3 binding cavity using combined in silico-in vitro approaches
Eda Suku

39 - Unraveling the histone code by fragment blind docking
Csaba Hetenyi

40 - Interacting residues prediction based on a random forest multi-step approach
R. Sanchez-Garcia

41 - Improved Rosetta protein structure prediction with customised fragments libraries based on structural class annotations
Jad Abbass

42 - PRODIGY: a structure-based method for the prediction of binding affinity in biomolecular complexes
Anna Vangone

43 - Folding membrane proteins by deep transfer learning
Jinbo Xu
44 - PERFORMANCE OF MACHINE-LEARNING SCORING FUNCTIONS IN STRUCTURE-BASED VIRTUAL SCREENING
Pedro Ballester

45 - Mining Functionally Conserved Building Blocks in Proteins
Florian Kaiser

46 - Unveiling the inhibition mechanism of HIF-2α:ARNT dimerization by protein dynamics
Stefano Motta

47 - Deep Learning strategy for Improving ranking of protein fold recognition method ORION
Jean-Christophe Gelly

48 - Molecular Dynamics Simulation of a 17β-Estradiol Specific DNA Aptamer
Alexander Eisold

49 - How is structural divergence related to evolutionary information?
Diego Javier Zea

50 - New binding site of the quorum sensing molecule N-3-Oxododecanoyl Homoserine Lactone with the transcriptional regulator LasR of Pseudomonas aeruginosa: Insights from Molecular Docking and Dynamics Simulations
Hovakim Grabski

51 - PhyrePower- Protein fold recognition by contact threading
Michael JE Sternberg

52 - Investigating the molecular determinants of ebolavirus pathogenicity
Mark Wass

53 - Sphinx: Merging Knowledge-Based and Ab Initio Approaches to Improve Protein Loop Prediction
Claire Marks

54 - Proteins From Peptides
Andrei N. Lupas

55 - Decoding the specificity of short linear motifs using spectrally encoded libraries: A curious case of Calcineurin
Nikhil P. Damle
56 - Homology modeling in a dynamical world  
Alexander Monzon

57 - Identifying Multiple Active Conformations of G Protein-Coupled Receptors Using Focused Conformational Sampling  
Ravinder Abrol

58 - Inferring protein phylogeny by modelling the evolution of secondary structure  
Jhih-Siang Lai

59 - What can human variation tell us about proteins?  
Stuart A. MacGowan

60 - New Insights into statistical potentials for describing protein binding affinity and aggregation properties  
Fabrizio Pucci

61 - Using Local States to Drive the Sampling of Global Conformations in Proteins  
Alessandro Pandini

62 - Understanding enterovirus uncoating by Normal Mode Analysis and Perturbation Response Scanning  
Caroline Jane Ross

63 - The determination of force field parameters of the conserved copper coordinating active site of AA9 proteins  
Vuyani Moses

64 - All-Atom Molecular Dynamics Simulations of a Membrane Protein Stabilizing β-sheet  
Maral Aminpour

65 - Determining Allosteric Hot Spots in Hsp70 using Perturbation Response Scanning  
David Penkler

66 - Improving fragment assembly protein structure prediction  
Charlotte Deane

67 - Prediction of differentially expressed genes from PCOS women by analyzing comorbidities associated with the disease  
Ashvini Desai
68 - Automated evaluation of quaternary structures from protein crystal structures
Jose Duarte

69 - Deep Learning in text mining for protein docking using full-text articles
Varsha D. Badal

70 - Protein Structures and their features in UniProtKB
Nidhi Tyagi

71 - The Impact of Conformational Entropy on the Accuracy of the Molecular Docking Software FlexAID in Binding Mode Prediction
Louis-Philippe Morency

72 - A novel computational framework to identify novel drug targets: a case study of kinase inhibitors
Hammad Naveed

73 - MESH-score a method for estimation of protein model Accuracy
Chen Keasar

74 - LiteMol suite: A comprehensive platform for fast delivery and visualization of macromolecular structure data
Radka Svobodova

75 - Automated Realization of RNA Structure from Interaction Topology
Matthew Wicker

76 - Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability
Abdullah Kahraman

77 - Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins
Wim Vranken

78 - Next generation structure-based antibody drug design with ABodyBuilder and PEARS.
Jinwoo Leem

79 - Predicting the total hydrophobic surface area of protein structures from sequence
Sanne Abeln
80 - Density-based clustering in structural bioinformatics: application to beta turns and antibody CDRs
Roland Dunbrack

81 - Scalable Data Analytics of the PDB with the MacroMolecular Transmission Format (MMTF) and Big Data Technologies
Peter W Rose

82 - Sensitive and efficient topology-independent structural alignment
Antonín Pavelka

83 - Interactome based drug design based on disease-disease relationships
Gaurav Chopra

84 - Dewetting the intracellular water pocket of the human connexin 26 hemichannel via a water-selective repulsive potential
Tomas Perez-Acle

85 - Mutations and Variations in Health and Disease: Protein Interaction Networks and 3D Structure
Franca Fraternali

86 - Accurate and reliable prediction of relative ligand binding potency in drug discovery - Applications in scaffold hopping transformations
Jianxin Duan

87 - Binding of cationic porphyrins to hemoglobin and cytochrome C by the method of molecular docking
Aram Gyulkhandanyan

88 - Protein S-palmitoylation site prediction using position-specific scores
Tatsuki Kikegawa

89 - Docking to homology models highlights the molecular determinants of ligand binding to the AhR
Sara Giani Tagliabue

90 - An exploration of the structural interactome of Rac1
Marijne Schijns

91 - Do trends in biomacromolecular structure quality inspire optimism?
Vladimír Horský
92 - Dexterity: A framework to use a smartphone as a 3D wand
Jenny Vuong

93 - Structural and functional analysis of alternative microexons of proteins observed in RNA-seq studies.
Matsuyuki Shirota

94 - A novel method for large-scale structural comparison of protein pockets using a reduced vector representation
Tsukasa Nakamura

95 - Inference of functional states from conformational changes in protein complexes
Markus Gruber

96 - Structural characterization of the IC pocket in the human Cx50 hemichannel
Claudia Pareja-Barrueco

97 - Using normal modes analysis to characterize the flexibility of protein tunnels and channels
Pierre Bedoucha

98 - Sensitive and efficient topology-independent structural alignment
Antonín Pavelka

99 - A Web Based 3D Visualization Solution for Large Scale Data
Chao-Chun Chuang

100 - LIBRA-WA: a web application for ligand binding site detection and protein function recognition
Fabio Polticelli

101 - Computational Approaches to Assessing Clinical Relevance of Pre-clinical Cancer Models
Vladimir Uzun
10 - Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone

Juan Rodriguez-Rivas¹, Simone Marsili¹, David Juan¹, Alfonso Valencia¹

¹Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, 28029 Madrid, Spain

Abstract

Protein–protein interactions are fundamental for the proper functioning of the cell. As a result, protein interaction surfaces are subject to strong evolutionary constraints. Recent developments have shown that residue coevolution provides accurate predictions of heterodimeric protein interfaces from sequence information. So far these approaches have been limited to the analysis of families of prokaryotic complexes for which large multiple sequence alignments of homologous sequences can be compiled. We explore the hypothesis that coevolution points to structurally conserved contacts at protein–protein interfaces, which can be reliably projected to homologous complexes with distantly related sequences. We introduce a domain-centered protocol to study the interplay between residue coevolution and structural conservation of protein–protein interfaces. We show that sequence-based coevolutionary analysis systematically identifies residue contacts at prokaryotic interfaces that are structurally conserved at the interface of their eukaryotic counterparts. In turn, this allows the prediction of conserved contacts at eukaryotic protein–protein interfaces with high confidence using solely mutational patterns extracted from prokaryotic genomes. Even in the context of high divergence in sequence (the twilight zone), where standard homology modeling of protein complexes is unreliable, our approach provides sequence-based accurate information about specific details of protein interactions at the residue level. Selected examples of the application of prokaryotic coevolutionary analysis to the prediction of eukaryotic interfaces further illustrate the potential of this approach.

References
11 - EncoMPASS: an Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry

Edoardo Sarti¹, Antoniya Aleksandrova¹, Lucy Forrest¹

¹National Institutes of Health

Abstract

EncoMPASS (Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry) is an online, completely automated database for relating integral proteins of known structure from the points of view of their sequence, structure, and symmetries. It can be used for organizing resources for protein structure determination, benchmarking sequence alignment tools, and inferring membrane protein functionalities via comparative studies.

Introduction

Integral membrane proteins constitute 20-30% of the genome and it has been estimated that they are targeted by around half of all FDA-approved drugs as well as of physiologically-relevant small ligands, making them extremely relevant in both cell biology and pharmacology. They are also associated with distinct structural features, such as the predisposition for internal and quaternary symmetries, that reflect the geometric constraints of the lipid bilayer. Several databases dedicated to structures of membrane proteins have been developed, but none of them classify the proteins or assign relationships between the proteins that they enumerate. Moreover, symmetry is never taken into consideration.

To address these issues, we present here the novel Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry (EncoMPASS), a fully-automated database through which we aim to introduce a more flexible representation of the structural relationships between experimentally-determined membrane protein structures.

Methods

In order to ensure the quality of our structural analysis, we select only proteins whose structure has been experimentally determined through X-ray crystallography with resolution <3.5 Å. Figure 1 illustrates the procedure for generating the data for one
database entry: the protein complex is first inserted in the membrane through the procedure used by the OPM database, then it is divided into single-chain subunits. Each chain is analyzed separately, and its sequence and structure are aligned with all other chains having the same topology. Both the complex and each individual chain then undergo symmetry recognition through two different programs, and the results thereof are combined in a consensus analysis.

**FIGURE 1.** EncoMPASS follows the depicted automatic protocol in order to perform all analyses for a given membrane protein complex and each of its single-chain subunits. Numerical and graphic data obtained from each of these analyses are available on the webpage describing that complex or chain.

Structure and sequence similarity are investigated using sequence identity, TM-score, RMSD, and through multiple graphical outputs that illustrate the distribution of related structures in the space of sequences and conformations, also at the residue level.

**Results & Conclusions**

EncoMPASS provides a structure- and symmetry-oriented analysis of over 2000 structures, covering more than 90% of all membrane protein coordinate files deposited in
the PDB. Data relating to sequence, structure and symmetry analyses is reported for each entry. A set of graphical tools helps the user to investigate how structures relate to each other in terms of sequence and structure similarity, and how frequently each residue of each single-chain subunit superposes in comparisons with other topologically similar chains.

A complete analysis of quaternary and internal symmetries is also available, and each symmetric region is illustrated and associated with the appropriate symmetry axis and symmetry transformation.

The database is updated monthly and its underlying code is freely available.

Thanks to these characteristics, EncoMPASS can be used for organizing resources for protein structure determination, benchmarking sequence alignment tools, and inferring membrane protein functionalities via comparative studies.

References

18 - Three-dimensional organisation of human genome

Przemysław Szałaj¹, Zhonghui Tang², Paul Michalski², Michał J Piętal¹, Michał Sadowski¹, Oscar Luo², Yijun Ruan², Dariusz Plewczynski¹

¹Centre of New Technologies, University of Warsaw, ²The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

Abstract

Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) reveals long-range chromatin interactions and provides insights into the basic principles of spatial genome organization and gene regulation mediated by specific protein factors. Recently, we showed that a single ChIA-PET experiment provides information at all genomic scales of interest, from the high resolution locations of binding sites and enriched chromatin interactions mediated by specific protein factors, to the low resolution of non-enriched interactions that reflect topological neighborhoods of higher-order chromosome folding. This multilevel nature of ChIA-PET data offers an opportunity to use multiscale 3D models to study structural-functional relationships at multiple length scales, but doing so requires a structural modeling platform.

Introduction

Chromosomal folding are important features of genome organization, which play critical roles in genome functions, including transcriptional regulation. Using 3C-based mapping technologies to render long-range chromatin interactions has started to reveal some basic principles of spatial genome organization. Among 3D genome mapping technologies, ChIA-PET is unique in its ability to generate multiple datasets (in a single experiment), including binding sites, enriched chromatin interactions (mediated by specific protein factors, like CTCF), as well as non-enriched interactions that reflect topological neighborhoods of higher-order associations. The multifarious nature of ChIA-PET data represents an important advantage in capturing multi-layer structural-functional information, but also imposes new challenges in multi-scale modeling of 3D genome [1].

Methods

The above experimental insights allowed us to propose [2] the 3D-GNOME (3-Dimensional GeNOme Modeling Engine), a complete computational pipeline for 3D
simulation using ChIA-PET data. 3D-GNOME consists of three integrated components: a graph-distance-based heatmap normalization tool, a 3D modeling platform, and an interactive 3D visualization tool. Using ChIA-PET and Hi-C data derived from human B-lymphocytes, we demonstrate the effectiveness of 3D-GNOME in building 3D genome models at multiple levels, including the entire genome, individual chromosomes, and specific segments at megabase (Mb) and kilobase (kb) resolutions of single average and ensemble structures. Further incorporation of CTCF-motif orientation and high-resolution looping patterns in 3D simulation provided additional reliability of potential biologically plausible topological structures.

**Results & Conclusions**

Finally, we will present 3D-GNOME web service [3], which generates 3D structures from 3C data and provides tools to visually inspect and annotate the resulting structures, in addition to a variety of statistical plots and heatmaps which characterize the selected genomic region. 3D-GNOME simulates the structure and provides a convenient user interface for further analysis. Alternatively, a user may generate structures using ChIA-PET data for the GM12878 cell line by simply specifying a genomic region of interest. 3D-GNOME is freely available at http://3dgnome.cent.uw.edu.pl/ providing unique insights in the topological mechanism of human variations and diseases.

Further refinement of 3DGNOME and application to additional ChIA-PET and other types of 3D genome mapping data will help to advance our understanding of genome structures and functions.

**References**


2. Computational model published in Genome Research 2016: http://genome.cshlp.org/content/early/2016/10/27/gr.205062.116.abstract

23 - From Mutations to Mechanisms and Dysfunction via Computation and Mining of Protein Energy Landscapes

Tatiana Maximova¹, Wanli Qiao¹, Erion Plaku², Carla Mattos³, Buyong Ma⁴, Ruth Nussinov⁴, Amarda Shehu¹

¹George Mason University, ²The Catholic University of America, ³Northeastern University, ⁴National Cancer Institute

Abstract

The energy landscape underscores the inherent nature of proteins as dynamic systems interconverting between structures with varying energies. Recently, we have developed a method that feasibly reconstructs landscapes. Here we demonstrate that the availability of landscapes of wildtype and diseased variants opens the way for data mining techniques to harness quantitative information embedded in landscapes to summarize mechanisms via which mutations alter dynamics and function.

Introduction

The energy landscape contains the information needed to characterize and relate protein equilibrium dynamics to function¹. Due to the broad spatio-temporal scales involved, neither wet- nor dry-laboratory techniques can reconstruct entire landscapes². In 2016 we proposed a method to do so for medium-size proteins by exploiting known structures of a protein's wildtype and variant sequences³. Our interest in understanding the impact of Ras mutations on (dys)function and now our ability to build landscapes for many variants prompt us to investigate landscape mining techniques. We report here on a technique that allows relating specific structures and transitions to biological mechanisms and exposing mechanisms via which mutations alter dynamics and function.

Methods

The SoPriMp algorithm³ we proposed and analyzed in 2016 leverages known structures of a protein to build a sample-based representation of the energy landscape of a sequence of interest. Due to the role of Ras in cell growth and of many of its variants in cancer and other disorders, we have applied SoPriMp to map landscapes of 14 pathogenic variants. Interesting questions can be asked on how mutations alter
dynamics and function via landscape comparison. We propose here a landscape mining technique based on statistical analysis of high-dimensional spatial data, borrowing concepts from manifold theory. The technique automatically extracts basins and saddles, allowing us to measure descriptors related to volumes, depths, and spatial and energetic distances between basins and basin-separating saddles.

Results & Conclusions

This technique yields graphical representations of computed landscapes, shown for selected variants in Figure 1A. On these (and others not shown) we observe: In oncogenic variants, the On-Off barrier typically gets elevated; the Off basin shrinks or disappears; the On basin splits, separating the R- and T-states (the importance of the latter for function is detailed in Ref. 4). On syndrome-causing variants, changes are less drastic: the Off basin leaks into other regions, or both the Off and On basins degenerate, even merging with one another and spilling over much of the landscape. Extracted landscape descriptors can be correlated with wet-lab measurements (on the same variants) of Ras activities (see Figure 1B). These and other descriptors (transition costs not shown here) reveal that the transition between R- and T-states relates to RAF1-RBD binding and GTP activation, and transitions between states with active-inactive GTP signaling (On-to-Off, T-state-to-Off) relate to intrinsic hydrolysis. These results signal an exciting stage where we can compute and mine landscapes to learn how mutations impact function and elucidate the role of specific states and transitions in biological pathways.
FIGURE 1. A. Landscapes of H-Ras variants. Contours track basins (left: On; right: Off). B. Cross-variant wet-lab-characterized mechanisms\textsuperscript{5} that correlate above 0.6 (in bold: above 0.75) with landscape descriptors are P7 (Intrinsic Hydrolysis) and P0 (GTP activation).

References

25 - How nature builds electrostatic interactions in natural enzymes: What can we learn for enzyme design?

Timothy A. Coulther¹, Lisa Ngu¹, Penny J. Beuning¹, Mary Jo Ondrechen¹

¹Northeastern University

Abstract

Enzyme design is in its infancy and basic design principles are not yet established. We show how strong electrostatic coupling between proton transfer equilibria is prevalent in the active sites of natural enzymes and facilitates catalysis. Multilayer networks of coupled charged residues promote catalytic efficiency, as observed in natural enzymes and in evolved artificial enzymes.

Introduction

Enzymes catalyze reactions at physiological temperature and neutral pH with high specificity. Therefore they hold great appeal as green industrial catalysts because of the potential for significant reduction in energy expenditures and in unwanted by-products. However, for most industrial chemical processes, no natural enzymes exist. Efforts to design enzymes to catalyze unnatural reactions have had some significant successes; indeed, stable folds can be designed in silico such that the necessary “reagent” residues are in the correct spatial arrangement to facilitate catalysis¹, ². These initial designs typically have very low activity; achieving significant activity requires many rounds of directed evolution over periods of years. The goal of this work is to understand the intrinsic properties that give natural enzymes their catalytic power and to learn how to build these properties into in silico enzyme design.

Methods

Partial Order Optimum Likelihood (POOL)³, ⁴ is a machine learning method developed by us to predict residues important for function, using the 3D structure of the query protein. The input features to POOL are based on computed electrostatic and chemical properties from THEMATICS. These input features are effectively measures of the strength of coupling between protonation events. POOL is used to characterize the properties of natural enzymes that are necessary for efficient catalysis and to explore how these properties may be built into designed enzymes.
Results & Conclusions

Catalytic sites in proteins are characterized by networks of strongly coupled protonation states; these networks impart the necessary electrostatic and proton-transfer properties to the active residues in the first layer around the reacting substrate molecule(s). Typically these networks include first-, second-, and sometimes third-layer residues. POOL-predicted, multi-layer active sites with significant participation by distal residues have been verified experimentally by single-point site-directed mutagenesis and kinetics assays for *Ps. putida* nitrile hydratase\(^5\), human phosphoglucose isomerase (Figure 1)\(^5\), *E. coli* Y family DNA polymerase DinB\(^6\), *E. coli* replicative DNA polymerase Pol III, and *E. coli* ornithine transcarbamoylase.

**FIGURE 1.** Multilayer active site of human phosphoglucose isomerase. Ligands in the substrate binding site are shown in space-filling form. First-layer residue side chains are colored green, second-layer blue, and third-layer orange.

In designed enzymes, such as retroaldolases, the residue-specific input features to POOL – measures of the strength of coupling between protonation equilibria – rise as the enzymes evolve to higher rates of catalytic turnover. An approach to build these properties into the initial designs is proposed.
References


28 - Computational design of a symmetrical beta-trefoil lectin with cancer cell binding activity

Daiki Terada¹,², Arnout R. D. Voet³, Hiroki Noguchi³, Kenichi Kamata¹, Mio Ohki¹, Christine Addy¹, Yuki Fujii⁴, Daiki Yamamoto⁵, Yasuhiro Ozeki⁵, Jeremy R. H. Tame¹, Kam Y. J. Zhang²

¹Graduate School of Medical Life Science, Yokohama City University, 1-7-29 Suehiro, Yokohama, Kanagawa 230-0045, Japan, ²Structural Bioinformatics Team, Division of Structural and Synthetic Biology, Center for Life Science Technologies, RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan, ³Laboratory of Biomolecular Modelling and Design, Department of Chemistry, KU Leuven, Celestijnenlaan 200G 3001, Heverlee, Belgium, ⁴Department of Pharmacy, Graduate School of Pharmaceutical Science, Nagasaki International University, 2825-7 Huis Ten Bosch, Sasebo, Nagasaki 859-3298, Japan, ⁵Laboratory of Glycobiology and Marine Biochemistry, Graduate School of NanoBio Sciences, Yokohama City University, 22-2, Seto, Yokohama, Kanagawa 236-0027, Japan

Abstract

Computational protein design has advanced very rapidly over the last decade, but there remain few examples of artificial proteins with direct medical applications. We describe a new artificial β-trefoil lectin that recognizes Burkitt’s lymphoma cells. The new protein, Mitsuba-1, contains 150-residues with three identical tandem repeats. Mitsuba-1 was expressed and crystallized to confirm the X-ray structure matches the predicted model. Mitsuba-1 recognizes cancer cells that express globotriose on the surface, but the cytotoxicity is abolished.

Introduction

MytiLec is a small lectin isolated from the Mediterranean mussel, and found to bind sugar chains with α-D-galactose. MytiLec-1 shows cytotoxic effects towards certain cancer cells including Burkitt’s lymphoma. Recently, we have created perfectly symmetrical proteins from natural templates based on the view that many nearly symmetrical ring-shaped proteins have evolved through exactly such an intermediate phase. We designed Pizza, a β-propeller protein with six identical blades, and showed it can fold readily and is extremely stable¹. We have further demonstrated that this symmetrical protein can be used as a template for biomineralization². Here we have adopted a similar procedure and applied it to MytiLec-1, to create a protein with three
identical subdomains, that retains sugar binding activity and the ability to bind selected cell types.

Methods

We used ancestral sequence reconstruction to derive likely parent sequences assuming evolution through duplication, and then computationally evaluated these sequences for stability. The detailed method is described as follows. A template structure for the desired symmetric protein to be designed is selected from the Protein Data Bank. All the individual subunits are structurally superimposed and their sequences aligned. A phylogenetic tree is created based on this alignment. Putative ancestral sequences are generated using a maximum likelihood based ancestral reconstruction algorithm. One representative subunit that is the closest to all the other subunits is identified and used to generate the backbone of an ideal protein with perfect symmetry. The energy of each one of the ancestral sequences adopting the ideal protein structure is calculated. One sequence with the lowest energy is selected and duplicated multiple times to create the protein with the desired number of identical sequences repeated in tandem. This sequence will be used to produce the recombinant protein for experimental verification.

Results & Conclusions

We have used this method to design a perfectly symmetric β–trefoil protein called Mitsuba-1 containing three identical subdomains. Each subdomain consists of 47 residues. Mitsuba-1 is cloned and expressed in E. coli and subsequently purified and crystallized. The crystal structure of Mitsuba-1 was determined at 1.54Å resolution. It confirmed the design and also revealed the binding of three GluNac ligands. Mitsuba-1 is stable with a melting temperature of 55°C determined by CD and remains folded in the presence of 3.6M GdmCl. Mitsuba-1 can bind to Raji cells, which are derived from Burkitt’s lymphoma. Mitsuba-1 is not found to reduce the viability of Raji cells. Mitsuba-1 is a further example demonstrating the effectiveness of our method for creating repeat proteins by examining probable evolutionary routes to existing natural proteins.
FIGURE 1. The overall structure of Mitsuba-1 viewed along the pseudo-three-fold symmetry axis or perpendicular to it.

References


42 - PRODIGY: a structure-based method for the prediction of binding affinity in biomolecular complexes

Anna Vangone¹, Li Xue¹, Rogridues Joao¹, Kastritis Panagiotis¹, Trellet Mikael¹, Shaarschmidt Jorg¹, Alexandre MJJ Bonvin¹

¹Utrecht University

Abstract

Here we present PRODIGY (PROtein bindIng enerGY predictor), a method to predict binding affinity of biomolecular complexes. Our approach is based purely on structural properties of the complexes, and outperforms other predictors so far reported. Due to its high performance and fast prediction time, we implemented it in a user-friendly web-server, freely available at: http://milou.science.uu.nl/services/PRODIGY/.

Introduction

Interactions between biomolecules, such as protein, nucleic acids and small ligands, are at the basis of almost every process happening in the cells. Understanding these interactions is therefore a crucial step in the investigation of biological systems and in drug development. Despite all efforts that have been devoted to unravel principles of biomolecular interactions in the past decades, we still lack a thorough understanding of the energetics of proteins association.

Recently, we introduced a simple but robust descriptor of binding affinity based only on structural properties of the protein-protein complexes (1). In this work, we have shown the contribution that the number and typology of the interfacial contacts made at protein-protein complex interface has in the description of the binding affinity, and developed a high performing contact-based predictor. Despite the importance of the topic, there are surprisingly only limited online tools for a fast and easy fast prediction of binding affinity. For this reason, based on our contact-based binding affinity predictor, we developed PRODIGY (PROtein bindIng enerGY predictor) (2), a webserver to predict the affinity of a protein-protein complex from its three-dimensional structure.

Methods
In order to evaluate the relationship between the contacts at the interface and the experimental binding affinity in protein-protein complexes, we used the bound structures from the protein-protein binding affinity benchmark in Kastritis et al. (3). We calculated the number of interface pair-wise contacts (ICs) for each complex (4) and the non-interacting surface properties (NIS) (5), classifying both ICs and NIS according to the charged/polar/apolar nature of the residue. We trained a linear regression model on such structural features (ICs and NIS) and assessed the linear dependency between the experimental binding affinity and the structural properties tested as Pearson’s correlation coefficient was reported. PRODIGY performance were compared with the current state of art by using the pre-calculated data available in CCharPPI (6).

Results & Conclusions

Using the protein-protein binding affinity benchmark of Kastritis et al. (3) we showed that the number of interfacial contacts (ICs) at the interface of a protein-protein complex correlates with the experimental binding affinity. This information, combined properties of the non-interacting surface (NIS) which have been shown to influence binding affinity, (5) has led one of the top binding affinity predictor in the field, showing Pearson’s Correlation $r = 0.73$, $p$-value $<$ 0.0001 and RMSE $=$ 1.89 kcal mol$^{-1}$ (Figure 1).
**FIGURE 1.** Scatter plot of predicted vs experimental binding affinities for a set of 81 protein-protein complexes.

Finally, we implemented our contact-based binding affinity predictor into the web-server PRODIGY (2).

**References**


43 - Folding membrane proteins by deep transfer learning

Jinbo Xu

1Toyota Technological Institute at Chicago

Abstract

Predicting membrane protein (MP) structures is challenging partially due to lack of sufficient solved structures for homology modeling. We describe a low-cost, high-throughput deep transfer learning method that first predicts MP contacts by learning from non-membrane proteins (non-MPs) and then predicts 3D structure models using predicted contacts. Tested on 510 non-redundant MPs, our method predicts correct folds for 218 of 510 MPs, of which 57 and 108 MPs having 3D models with RMSD less than 4Å and 5Å, respectively. Rigorous blind test in CAMEO shows that our method predicted correct folds for all 4 test MPs and high-resolution 3D models (RMSD ~ 2Å) for two. We estimated that our method could predict correct folds for 1345~1871 of 2215 reviewed human multi-pass MPs, including a few hundred new folds.

Methods

We predict protein contacts by concatenating two deep residual neural networks (see [1, 2] for details), which performed the best in CASP12 (for soluble protein contact prediction). We extend this method to MP contact prediction. Compared to soluble protein contact prediction, one challenge of applying deep learning to MP contact prediction is lack of sufficient training data since there are only 510 non-redundant MPs with solved structures in PDB. To overcome this, we train our deep learning model using thousands of non-MPs with solved structures. It turns out that the resultant deep model works well for MP contact prediction. Our further study indicates that using a mix of non-MPs and MPs with solved structures, we can train a deep model with even better prediction accuracy, greatly outperforming existing methods. Our method is of low cost and very high throughput. It takes from 30 minutes to a few hours on a Linux workstation to predict 3D models for a test MP, much more efficient than the popular fragment assembly methods.

Results
**Contact prediction accuracy.** Tested on the 510 non-redundant MPs, when top L/2 predicted long-range contacts are evaluated, our method has accuracy 0.58, much better than pure co-evolution methods such as Evfold (0.23), PSICOV (0.25) and CCMpred (0.28), and a supervised learning method MetaPSICOV (0.37). This result suggests that non-MPs and MPs share some common properties for contact prediction that can be learned by our deep learning model, and that the set of non-MPs contain more information for contact prediction than the set of MPs.

**Folding accuracy.** Experimental results show that our predicted contacts can help fold about 218 and 288 of the 510 MPs with solved structures when TMscore≥0.6 and TMscore≥0.5 are used as the correctness criterion, respectively. Meanwhile, 57 and 169 of the 510 MPs have predicted 3D models with RMSD<4.0Å and <6.5Å, respectively. In contrast, pure co-evolution methods such as CCMpred (and PSICOV, Evfold) and homology modeling can fold only 10 and 18 of the MPs with RMSD<4.0Å, respectively, and 49 and 89 of them with RMSD<6.5Å.

**Blind test in CAMEO.** We have implemented our algorithm as a fully-automated web server and blindly tested it through the weekly live benchmark CAMEO (http://www.cameo3d.org/) operated by Schwede group. CAMEO can be interpreted as a fully-automated CASP, with >30 participating servers including some well-known such as Baker’s Robetta, RaptorX, Swiss-Model and HHpred. Since September 2016, there are only four test MPs among all the CAMEO hard targets. Our web server (CAMEO ID: Server60) predicted correct folds for all of them, much better than all the other CAMEO-participating servers. In particular, for 5h36A and 5h35E (>210 AAs) we predicted 3D models with RMSD ~ 2Å.

**References**


54 - Proteins From Peptides

Andrei N. Lupas¹, Vikram Alva¹, Marcus D. Hartmann¹, Edgardo Sepulveda¹, Jörg Martin¹, Hongbo Zhu¹

¹Max Planck Institute for Developmental Biology, Tübingen, Germany

Introduction

For the most part, contemporary proteins can be traced back to a basic set of a few thousand domain prototypes, many of which were already established in the Last Universal Common Ancestor of life on Earth, around 3.5 billion years ago. The origin of these domain prototypes, however, remains poorly understood. We have proposed that they arose from an ancestral set of peptides, which acted as cofactors of RNA-mediated catalysis and replication¹. Initially, these peptides were entirely dependent on the RNA scaffold for their structure, but as their complexity increased, they became able to form structures by excluding water through hydrophobic contacts, making them independent of the RNA scaffold. Their ability to fold was thus an emergent property of peptide-RNA coevolution.

The ribosome is the main survivor of this primordial RNA world and offers an excellent model system for retracing the steps that led to the folded proteins of today, due to its very slow rate of change². Close to the peptidyl transferase center, which is the oldest part of the ribosome, proteins are extended and largely devoid of secondary structure; further from the center, their secondary structure content increases and supersecondary topologies become common, although the proteins still largely lack a hydrophobic core; at the ribosomal periphery, supersecondary structures coalesce around hydrophobic cores, forming folds that resemble those seen in proteins of the cytosol. Collectively, ribosomal proteins chart a path of progressive emancipation from the RNA scaffold, offering a window onto the time when proteins were acquiring the ability to fold.

Results & Conclusions

We have retraced this emancipation from the RNA scaffold computationally and experimentally for a cytosolic protein fold, the tetratricopeptide repeat (TPR). By amplifying an α-hairpin from a ribosomal protein, RPS20, which is unstructured in the absence of the cognate ribosomal RNA, we explored whether an intrinsically disordered peptide could form a folded protein through an increase in complexity afforded by
repetition. Simple repetition was not sufficient in our case, but the repeat protein was so close to a folded structure that only two point mutations per repeat were necessary to allow it to fold reliably. The mutations needed for this transition did not appear to affect negatively the interaction with the RNA scaffold and were neutral for survival and growth in the parent organism, raising the possibility that they could have been among the variants sampled multiply in the course of evolution. TPRs could thus have plausibly arisen by amplification from an ancestral, RNA-dependent helical hairpin, as proposed by our theory.

**FIGURE 1.** (A) Scenario for the divergent evolution of ribosomal protein RPS20 and the cytosolic TPR fold from an ancestral, ribosome-associated α-hairpin. (B) The crystal structure of the RPS20-derived repeat. The three chains in the asymmetric unit are
colored green, blue and yellow, respectively. Chains A and B form a dimer. (C) Superposition of the RPS20-derived repeat (green) and the TPR protein CTPR3 (PDB: 1na0, chain A, gray).

References


Identifying Multiple Active Conformations of G Protein-Coupled Receptors Using Focused Conformational Sampling

Ravinder Abrol¹, Sijia Dong², William A. Goddard III³

¹California State University Northridge, ²University of Minnesota, ³California Institute of Technology

Abstract

G protein-coupled receptors (GPCRs) are membrane proteins critical in many cellular signal transductions. The pleiotropic signaling of GPCRs is enabled by their conformational flexibility that enables them to exist in multiple states, where functionally important active states are high in energy. This makes the experimental studies of these active states very challenging. Most computational methods can only identify lowest-energy states, so we developed a focused conformational sampling method that is capable of identifying multiple active states of GPCRs. It was able to correctly predict the active conformation of two GPCRs starting only from the inactive state, and explained previous experiments, which has been an unsurmountable challenge for standard molecular dynamics simulations.

Introduction

The multiple functions of GPCRs are dependent on their activation, which enables them to couple to several signaling partners like G proteins, GPCR kinases (GRKs), and arrestins inside the cell. This has also led to the pharmacological targeting of GPCRs that causes both therapeutic effects and undesirable side-effects. A mechanistic understanding of GPCR functions requires the study of multiple active conformations that these receptors can putatively reside in. These conformations have been hard to characterize experimentally as they are high in energy and difficult to stabilize. The computational biophysical methods are good at identifying mainly lowest energy conformations. To address this problem, we have developed the ActiveGENSeMBLE computational method¹ that systematically predicts multiple conformations that are likely in the GPCR activation landscape, including multiple active- and inactive-state conformations, while minimizing computationally cost.
Method

ActiveGEnSeMBLE method\(^1\) starts with a systematic coarse grid conformational sampling of helix tilts/rotations (~13 trillion transmembrane (TM)-domain conformations) and selects the conformational landscape based on energy. This profile (Figure 1) identifies multiple potential active-state energy wells, using the TM3-TM6 intracellular distance as a surrogate activation coordinate. These energy wells are then sampled locally using a finer grid in conformational space to find a locally minimized conformation in each energy well (Figure 1), which can be further relaxed if necessary using molecular dynamics (MD) simulations.

![Figure 1: Conformational sampling in ActiveGEnSeMBLE method.](image)

Results and Conclusions

We validated the ActiveGEnSeMBLE method by predicting active human β\(_2\) adrenergic receptor (hβ2AR) and human M\(_2\) muscarinic acetylcholine receptor (hM\(_2\)R) crystal structures. We found that the ActiveGEnSeMBLE method sampled the orientations of the TM helices and located structures in various energy wells spanning the range of TM3–TM6 distances (R\(_{36}\)) traversed in the process of activation. Subsequent analysis revealed a local minimum in each of these energy wells that was close or identical to a crystal-structure conformation based on backbone RMSD (Figure 2).
To show the utility of the knowledge of active conformations, molecular dynamics simulations of the active conformation of hβ2AR, with and without the G protein and the agonist, were used to generate energy profiles that are consistent with the qualitative energy landscape of hβ2AR obtained from experiments\textsuperscript{2}, providing information about how the ligand and G protein may play a role in activation. These results (Figure 3) indicate that the agonist alone is not enough to stabilize the active state, and that the G\textsubscript{α} C-terminal chain needs to be bound to the GPCR to promote activation, in agreement with experimental observations\textsuperscript{2}.

---

**Figure 2:** Conformational energy landscape for human β2 adrenergic receptor activation.
of human β2 adrenergic receptor.

References


What can human variation tell us about proteins?

Human sequencing projects have generated population variant datasets from thousands of individuals. For the exome, protein structure is an effective context in which to interpret the effects of missense variation. Studies have shown that disease variants are enriched in buried sites and protein interaction surfaces, whilst somatic and pathogenic germline variants often cluster in 3D. Beyond structure, the genomic distribution of genetic variation is affected by gene essentiality, protein domain architecture and other genomic features.

Given these observations, we were curious to see if missense variant distributions could identify the structural features that had been used to interpret them. The variant data were too sparse to compare genetic variation between individual protein residues in domains directly, so we began by aggregating variants over columns in multiple sequence alignments of Pfam domains. We then compared the resulting population variation profiles to residue conservation across all species and found that the same structural and functional pressures that affect residue conservation during domain evolution also place constraints on missense variants. Consequently, positions that are depleted of missense variants are likely to be structurally important.

We found that missense depleted positions are enriched in known pathogenic variants while positions that are both missense depleted and evolutionary conserved are further enriched in pathogenic variants compared to those that are only evolutionary conserved. Unexpectedly, some evolutionary Unconserved positions are Missense Depleted in human (UMD positions) and these are also enriched in pathogenic variants. UMD positions are further differentiated from other unconserved residues in that they are enriched in ligand, DNA and protein binding interactions, which suggests this stratification can identify unconserved positions that are structurally important.
I will illustrate these principles by discussing examples from the Src Homology 2 (SH2; Figure 1), G-Protein Coupled Receptor (GPCR; Figure 2) and the Nuclear Receptor Ligand Binding Domain families.

**Figure 1.** Inter-domain interactions of the SH2 domain in inactivated Src (PDB ID: 2src). The surface of the SH2 domain (PF00017) is coloured yellow to red corresponding to a. increasing missense depletion and b. increasing Shenkin conservation.

**Figure 2.** UMD residues (blue) in the Rhodopsin-like receptors (PF00001) mapped to a structure of the Delta-type
opioid receptor (PDB ID: 4n6h). Amongst the 11 UMD residues are several involved in ligand binding and one that coordinates the bound sodium ion.

References

66 - Improving fragment assembly protein structure prediction

Charlotte Deane¹, saulo de Oliveira¹

¹University of Oxford

Abstract

Protein structures can elucidate functional understanding, explain disease mechanisms and inform drug design. However, experimental structure determination is costly, and technically difficult and while the three-dimensional structure of proteins is difficult to obtain amino acid sequences are easily available and far outnumber solved structures. However, current de novo protein structure prediction methods are heuristics limited by the enormous search space, with successful prediction largely restricted to small, single domain proteins.

The three key components of de novo fragment-assembly methods for protein structure prediction are the fragment library, the “energy” function and the search method. In this talk I will give an overview of my groups work on improving each of these stages. Firstly, describing the development of a novel fragment library Flib that uses predicted secondary structure to determine library generation strategy [1]. Secondly, giving a comparison of the different co-evolution contact predictors in terms of their ability to improve protein structure prediction [2]. Finally demonstrating how sequential prediction approaches using SAINT2 can improve both search heuristics and final model quality.
The precision of five fragment library generators showing the differences in precision achieved for different secondary structure types.

References

1. Saulo Henrique Pires de Oliveira, Jiye Shi, Charlotte M Deane, Building a better fragment library for de novo protein structure prediction, Plos One, 2015, 10(4), e0123998

68 - Automated evaluation of quaternary structures from protein crystal structures

Jose Duarte¹, Spencer Bliven²,³, Aleix Lafita², Guido Capitani², Stephen K. Burley¹,⁴

¹RCSB Protein Data Bank, UCSD, ²Paul Scherrer Institute, ³National Center for Biotechnology Information, NIH, ⁴RCSB Protein Data Bank, Rutgers University

Introduction

Crystallography is the most powerful technique for generating atomic level structures of proteins and other biological macromolecules. However, it does not always yield definitive insights into the quaternary structures of biological macromolecules. In order to provide better tools for determining the most likely quaternary structure in proteins, we have developed the new EPPIC 3 method. It uses evolutionary considerations as the ultimate arbiters of the biological relevance of interfaces and assemblies, thereby offering a complementary approach versus other available methods that rely on thermodynamic considerations.

Results & Conclusions

EPPIC 3 extends our previous Evolutionary Protein-Protein Interface Classifier (EPPIC)¹, by going beyond classifying pairwise interfaces. It identifies all possible topologically valid assemblies present in a protein crystal and provides predictions as to likely quaternary structures.

Pairwise interface classifications are based on two evolutionary scores and a single geometric score. These descriptors were trained against a large dataset of known biologically relevant and crystal interfaces to fit a logistic regression classifier that provides probabilistic scoring of interfaces together with confidence assignment.

Assembly enumeration is achieved by representing the crystal lattice as a periodic graph. Finding valid assemblies is then reduced to the problem of finding subgraphs complying to a set of rules, which guarantees closed assemblies (Point Group symmetries) and isomorphism in the assembly composition and connectivity throughout the crystal. Finally the assemblies are scored based on the individual scores of the
constitutive interfaces, providing in the end a single probability of an assembly being the biological one, together with a confidence estimation. The confidence values are very valuable not only for users but also for downstream analyses (e.g. docking), where only high confident predictions can be selected.

The software is accessible through an easy to use web graphical interface at http://www.eppic-web.org. The graphical interface is designed to aid the crystallographer in interpreting putative quaternary structures using 2D and 3D graphical tools that operate within the browser (Figure 1).

The server offers additional useful tools to the structural bioinformatics community such as precalculated sequence alignments for every PDB structure and visualization of conservation on protein surface with the browser embedded NGL viewer. All the data are provided via xml downloads to the community, enabling further analyses.

**FIGURE 1.** Novel visualization tools are available as part of the web user interface, allowing the visualization of the crystal lattice via the lattice graph representation. This is done both in 2D thanks to the vis.js library and in 3D thanks to the fast NGL molecular visualization package\(^3\). The colors and labels show the different protein entities and interface types present in the lattice, allowing for a one-glance understanding of the assembly symmetry and its connectivity.

References
69 - Deep Learning in text mining for protein docking using full-text articles

Varsha D. Badal¹, Petras J. Kundrotas¹, and Ilya A. Vakser¹

¹Center for Computational Biology, The University of Kansas, Lawrence, Kansas 66047, USA

Residues extracted from PubMed abstracts by text mining techniques can be used as constraints in protein-protein docking (Badal, et al., 2015). However, the pool of the mined residues contains many false positives (residues not relevant to protein binding), which can be partially removed by natural language processing algorithms (Badal et al, 2017, submitted). Deep learning methods can potentially provide further reduction of these false positives. However, abstracts provide more formal and strictly crafted text, which may lack in variety/richness for training of the deep learning models. On the other hand, full text articles, while providing richer source of information, are freely available only for a fraction of the PubMed abstracts, as PMC-open access. We investigated whether deep-learning models trained on the limited-availability full texts can be applied for the filtering of residues in the PubMed abstracts. For this purpose, we used deep recursive neural network, which composes word vectors (Irsoy and Cardie, 2014). Word vectors of the residue-containing sentences from the PMC full text articles were generated by word2vec (Mikolov, et al., 2013). We propose to label words and trees with sentiments from 0 to 4 (0,1,2 labeling negative samples, i.e. describing non-interface residues and 3,4 denoting positive samples, relevant to the interface residues) to train, and subsequently, classify sentences containing residues. The approach was tested on the non-redundant set of protein complexes from DOCKGROUND resource (http://dockground.compbio.ku.edu). The results showed that the model is capable of distinguishing the abstract sentences containing interface residues from those containing non-interface residues. We further investigated the local sentiment (surrounding words/phrase) using a window of words of various lengths around the residue in a sentence.

References


The Impact of Conformational Entropy on the Accuracy of the Molecular Docking Software FlexAID in Binding Mode Prediction

Louis-Philippe Morency\textsuperscript{1,2}, Rafael Najmanovich\textsuperscript{1,2}

\textsuperscript{1}Université de Montréal, \textsuperscript{2}PROTEO

Abstract

Here we show the newest implementation of Flexible Artificial Intelligence Docking (FlexAID) allowing its scoring function to consider the conformational entropy of ligand and biomolecules complexes. The higher accuracy of FlexAID\textsuperscript{1} on complex cases, the addition of novel features, i.e. the conformational entropy, its accessibility and its easy-to-use graphical user interface place FlexAID in an interesting position to tackle biologically and pharmacologically relevant situations currently ignored by other methods.

FlexAID\textsuperscript{1} is available as a command-line pre-compiled executable (available at http://bcb.med.usherbrooke.ca/flexaid for Windows, macOS & Linux) or through the NRGsuite, a PyMOL integrated user interface allowing the user to use FlexAID in an intuitive manner with real time visualization. Both the NRGsuite\textsuperscript{2} and FlexAID\textsuperscript{1} are distributed as open-source software.

Introduction

A major goal of molecular docking is to predict the experimentally observed binding mode between a biomolecule, i.e. a polymer of amino or nucleic acids, and a ligand—e.g. small molecules, peptides or nucleic acids. This computational method is used to study the structure of the molecular interactions involved in cell's essential biological functions. Actual molecular docking methods are developed to evaluate the molecular interactions of a single conformation, a single pose at a time, and they are trained to estimate the enthalpic fraction of the binding free energy. Consequently, most current molecular docking methods fail to efficiently model the entropic contributions, especially those of conformational nature, who are fundamental in molecular recognition events.
Our research group develops FlexAID\(^1\), an accessible and competitive ligand and biomolecule molecular docking software whose focus is on molecular flexibility. Here we introduce FlexAID’s newest feature that allows its scoring function to estimate the conformational entropy by redefining the static binding mode usually predicted in molecular docking into a dynamic collection of similar poses evaluated altogether.

**Methods**

We implemented the new scoring function in FlexAID, allowing its genetic algorithm to select less favourable, but frequently observed binding modes, with a probability following a Boltzmann distribution. This implementation allows FlexAID to consider conformational entropy of the complexes during the molecular docking simulation. The core of the implementation of the conformational entropy in FlexAID resides within an unsupervised and density-based molecular classification algorithm charged to group similar poses together into binding modes, thus redefining a binding mode as a dynamic collection of poses scored altogether as it an be seen in Figure 1.

![Figure 1. Visual representation of a dynamic binding mode output by the newest implementation of FlexAID (PDB: 1xm6).](image)

**Results & Conclusions**

We present the impact of FlexAID’s newest feature on its accuracy in binding mode prediction using three increasingly complex scenarios: the Astex Diverse Set\(^4\), the Astex Non Native Set\(^5\) and the HAP2\(^6\) dataset. We show that FlexAID outperforms other open-source molecular docking methods when molecular flexibility is crucial. Furthermore, FlexAID now outputs multiple conformations per binding mode, a novelty that allows the user to visualize the dynamics of the complex studied. We believe that its higher accuracy in complex scenarios, the addition of novel features, e.g. the conformational
entropy, its accessibility and its easy-to-use graphical user interface, the NRGsuite\textsuperscript{2}, place FlexAID in an interesting position to tackle biologically and pharmacologically relevant situations currently ignored by other molecular docking methods.

References

73 - MESH-score a method for estimation of protein model Accuracy

Chen Keasar\(^1\), Tomer Sidi\(^1\)

\(^1\)Department of Computer Science, Ben-Gurion University, Israel

Introduction

The first steps of protein structure prediction (PSP) methods generate many alternative 3D models of the protein (aka decoys), and mediocre decoys typically outnumber the good ones. Thus the next step, an estimation of model accuracies (EMA, aka QA), has considerable impact on the overall PSP performance. EMA scores are also important for PSP users, as estimates of models' reliability.

Over the last two decades, we and others have developed quite a few energy terms and other structural features that provide different perspectives on the complex structures of proteins. Each feature assesses some aspect of the decoys’ quality, and EMA is the art of combining them into a single coherent score. The complexity of protein structures suggests that many diverse features combined in subtle ways may be needed to create a good EMA. Yet most EMA methods make do with relatively few, carefully selected, features to avoid the risk of overfitting. We take the opposite approach, ultimately aiming to make use of any structural feature we deem useful, and to combine these features in complex ways. To this end we developed an EMA platform called MESH-score\(^1\). We continuously augment it with more features and improve their integration. Here we present the unique approach, its performance in CASP12, and more recent results.

Methods

Decoy preprocessing: Before assessment, MESH-score reduces noise by side-chain repacking\(^2\) and energy minimization.

Ensemble learning and feature selection: MESH-score pursues an ensemble learning approach that combines many (usually 1000) independently trained predictors. Each predictor is assigned a somewhat different objective function, and is trained by stochastic optimization, which includes feature selection. Overfitting at the predictor level
is prevented by a constraint on the number of features with non-zero weight. The final score is a weighted median of the predictors' scores. Thus, each of the many features "has a chance" to contribute and yet, since the integration step does not include any adjustable parameters, overfitting is avoided.

Additionally, in a variant of MESHIScore, called MESHIScore-con, the decoys' score is biased towards the weighted average of its close neighbors (GDT_Ts >= 0.95) scores.

Features – Currently feature set of MESHIScore includes 106 diverse features, including pairwise potentials, torsion angle and hydrogen bond terms, contact and radius of gyration terms, compatibility with sequence based predictions, and meta-features that evaluate the distribution of other features.

Results

In CASP12, MESHIScore and MESHIScore-con participated as EMA servers. They were ranked among the top four or five top methods (and the best servers) according to two measures of performance (Fig. 1).

In a different CASP track, MESHIScore was also used by five structure prediction groups and three of them attained top ranks (Fig. 2).

The development of MESHIScore is a continuous process of feature incorporation and improved training. Table 1 depicts the performance improvement of MESHIScore over the last two years. It also compares MESHIScore with three other alternative scores.

Summary

MESHIScore is a modular and extendable EMA method with state-of-art performance.

References


A. http://www.predictioncenter.org/casp12/qa_diff2best_plots.cgi


**FIGURE 1.** Two screenshots from the CASP site, each depicts the top ranking groups by a different measure.5. Ray, Lindahl, and Wallner BMC bioinformatics 13:224 (2012)
FIGURE 2. Top structure prediction groups from the CASP12 site. Groups that uses MESHI-score are marked by arrows.

76 - Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability

Abdullah Kahraman¹,³, Pascal Leuenberger², Stefan Ganscha², Valentina Cappelletti², Paul J. Boersema², Christian von Mering¹, Manfred Claassen², Paola Picotti²

¹University of Zurich, ²ETH Zurich, ³Sabanci University, Istanbul, Turkey

Abstract

Temperature-induced cell death is thought to be due to protein denaturation, but the determinants of thermal sensitivity of proteomes remain largely uncharacterized. We developed a structural proteomic strategy to measure protein thermostability on a proteome-wide scale and with domain-level resolution. We applied it to E.coli, S.cerevisiae, T.thermophilus, and human cells, yielding thermostability data for more than 8000 proteins. Our results (1) indicate that temperature-induced cellular collapse is due to the loss of a subset of proteins with key functions, (2) shed light on the evolutionary conservation of protein and domain stability, and (3) suggest that natively disordered proteins in a cell are less prevalent than predicted and (4) that highly expressed proteins are stable because they are designed to tolerate translational errors that would lead to the accumulation of toxic misfolded species.

Introduction

Temperature is crucially important to life. Small temperature changes can differentiate optimal and lethal growth conditions of living organisms. Because of the higher abundance and lower stability of proteins as compared with those of other biological macromolecules, thermally induced cell death is thought to be due to protein denaturation, but the determinants of thermal sensitivity of proteomes remain largely uncharacterized.

Methods

To determine the thermal stability of proteins on a proteome-wide scale and with domain-level resolution, we developed a structural proteomic approach that relies on limited proteolysis (LiP) and mass spectrometry (MS) applied over a range of temperatures.
Results

Our LiP-MS strategy was validated through analysis of purified proteins in the presence and absence of a biologically relevant matrix. We then obtained proteome-wide thermal denaturation profiles for *E.coli*, *S.cerevisiae*, *T.thermophilus*, and human cells. In contrast to previous predications that proteome instability derives from the simultaneous and generalized loss of hundreds of proteins, we observed that at a temperature at which cells experience temperature-induced physiological impairment, a subset of essential proteins undergoes denaturation.

Confirming results of previous studies on the basis of comparison of genomes of thermophilic and mesophilic bacteria, we observed enrichment for lysine residues and b-sheet structures in thermostable proteins. We also found that unstable proteins have a higher content of aspartic acid than that of stable proteins and observed an inverse correlation between protein length and thermal stability. Further, thermostable proteins are substantially less prone to thermal aggregation than unstable proteins. Relative domain thermostability was conserved both within species and across organisms. Thermal stability was not generally similar for proteins encoded by orthologous genes. This suggests that the melting temperatures of proteins are affected by the reshuffling of protein domains, despite the conservation of domain stability.

According to the “translational robustness” theory, highly expressed proteins must tolerate translational errors that can lead to the accumulation of toxic misfolded species. Our data show a clear direct relationship between protein thermal stability and intracellular abundance and an inverse relationship between protein stability and aggregation or local unfolding. Increasing the thermodynamic stabilities of the folds of abundant proteins will broaden the range of amino acid replacements that a protein can tolerate before misfolding. Our findings suggest that over the course of evolution, the burden of intracellular misfolding has been reduced by increasing the thermodynamic stability of abundant proteins.

Conclusions

Our study contributes insight into the molecular and evolutionary bases of protein and proteome thermostability and provides a blueprint for future studies on the stability of proteomes and thermal denaturation.
77 - Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins

Daniele Raimondi\textsuperscript{1,2,3}, Gabriele Orlando\textsuperscript{1,2,3}, Rita Pancsa\textsuperscript{4}, Taushif Khan\textsuperscript{1,2}, Wim Vranken\textsuperscript{1,2}

\textsuperscript{1}Interuniversity Institute of Bioinformatics in Brussels, ULB/VUB, Triomflaan, BC building, 6th floor, CP 263, 1050 Brussels, Belgium, \textsuperscript{2}Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium, \textsuperscript{3}Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe, CP 212, 1050 Brussels, Belgium, \textsuperscript{4}MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, United Kingdom.

Abstract
The very early stages of protein folding, defined by intrinsic local interactions between amino acids close to each other in the protein sequence, are poorly understood. We developed a method to predict from the primary sequence of a protein where early folding is likely to take place, and show that these early folding predictions give insights into the folding process. On a proteome scale the predicted early folding residues tend to become the residues that interact the most in the folded structure, and are often the residues that display evolutionary covariation. These connections suggest that the initial behavior of the protein chain has a lasting effect on its subsequent states.

Introduction
Amino acids close to each other in the sequence that interact favorably are important in shaping the folding landscape, but the structure of a protein does not necessarily provide direct information about where the first local structural elements started to form. We created the Start2Fold database that identifies these residues from pulsed labelling and related Hydrogen Deuterium eXchange (HDX) experiments\textsuperscript{1}. Based on an additional analysis in relation to protein backbone rigidity predictions\textsuperscript{2}, we developed EFoldMine, a predictor of early folding residues based on data from high-quality experimental NMR-based HDX studies.

Methods
From 30 Start2Fold sequences covering a total of 3398 residues\textsuperscript{2}, 482 were designated ‘early folding’: these are the first residues to be involved in sufficient local structure...
formation to protect their backbone HN from solvent. We used 5 features computed from the protein sequence: DynaMine backbone dynamics\textsuperscript{3,4}, and a new set of predictions for side-chain dynamics and secondary structure propensity from NMR chemical shift-based estimations of the side-chain dynamics and the secondary structure propensity. We used a window of flanking residues between i-2 and i+2, resulting in a 25-dimensional feature vector. Performances were evaluated in strict stratified cross-validation settings, using BLASTCLUST to stratify the 30 sequences in function of a 25\% sequence identity (SI) cutoff at 90\% coverage.

**Results & Conclusions**

The EFoldMine performance reaches an MCC of 35.4, and an AUC of 80.8. We investigated how the EFoldMine predictions relate to two protein pairs that have a very similar topology but different folding pathways. EFoldMine picks up where folding starts, with very different early folding profiles between the pairs of proteins (Figure 1). The predictions also relate very well to independent HDX-MS experiments. On a proteome scale, they consistently encompass many of the residues that form the most extensive contacts in the folded protein, and not only for the typical (hydrophobic) structure-forming residues. We also observe that residues with evolutionary covariance signals tend to be early folding residues. The combination of these observations suggests that in particular the interactions between early folding residues have to be conserved in order to maintain the protein fold in evolution.

EFoldMine identifies the amino acid residues in proteins that are inclined to form structural elements unaided at the very first stage of the folding process. The connection of the early folding predictions to both folding pathway data and the folded protein structure suggests that the initial statistical behaviour of the protein chain has a lasting effect on its subsequent states.
FIGURE 1. Myoglobin (a,b) and leghemoglobin (c,d) in relation to early folding scores.

References


80 - Density-based clustering in structural bioinformatics: application to beta turns and antibody CDRs

Maxim Shapovalov¹, Benjamin North², Simon Kelow¹, Roland Dunbrack¹,²

¹Fox Chase Cancer Center, ²Temple University

Introduction

Clustering methods can be classified into a number of different categories: centroid-based clustering (e.g., k-means); connectivity-based clustering (e.g., hierarchical clustering); distribution-based clustering (e.g., Gaussian mixture models); and density-based clustering, which looks for local maxima in the data density separated by low-density regions. In this work, we explore the utility of density-based clustering methods in structural bioinformatics. An important advantage of density-based methods is that typically they identify outliers during the clustering process, and the outliers are not included in the output clusters. This may occur in protein structures when an element of structure (e.g., a protein loop) may be incorrectly modeled, or might be the result of human engineering rather than a natural process (e.g., antibody CDRs). It is of value to identify both the clusters of data and the noise, and both may be the subject of further study.

We have utilized "density-based spatial clustering of applications with noise" (DBSCAN) and its variants in two important applications. DBSCAN is a simple algorithm in which every data point is assigned to one of three categories, depending on the number of neighbors it has within a predefined distance, Eps. If a data point has at least a fixed number of points (defined by a parameter MinPts) within Eps, it is called a “core point.” Any point that is not a core point but is within Eps of a core point is a “border point.” All other points are called “noise points.” Core points are connected to each other (in the sense of a graph edge) if they are within Eps of each other. Each connected graph of core points becomes a cluster. Border points are assigned to same cluster as their nearest core point. Noise points remain unclustered.

Results
β turns are typically defined as non-helical, 4-residue segments in which the Cα atom of residue 1 is ≤ 7 Å from the Cα atom of residue 4. Turns have been analyzed since the 1970s, and the extant β turn types comprise the following: I, I’, II, II’, VIa(1,2), VIb, and VIII. The Type VI turns have a cis-peptide bond between residues 2 and 3. Type IV turns are a catch-all class for turns that do not fit the standard types, and comprise up to a third of β turns. We compiled a set of 13,030 β turns from 1082 protein chains from structures with resolution ≤ 1.2 Å and less than 50% mutual sequence identity. We used a dihedral angle metric, which is a sum over the six dihedral angles connecting Cα(1) to Cα(4) of the function $D_{ij}=2(1-\cos(x_i - x_j))$ for each dihedral $x$. DBSCAN was successful at removing 289 noise points (2.2%) and produced 10 clusters of β turns, 3 of which can be subclustered with a further round of DBSCAN with different parameters or with another clustering method. From these results, we have developed a new, simpler nomenclature for β turns from these (preliminary) results based on the Ramachandran regions: A (for α), B (for β and polyproline II regions), L (for α-left), and E (for the lower right and upper right regions) and lower case for cis residues. In addition to the standard turns (Types AA, LL, BL, EA, Ba, Bb, AB respectively), we find small but significant populations of AL and LA turns (shown below) and two turn types with a cis peptide bound before residue 2: aA and aB.

We have used DBSCAN and an Linfinity metric (the maximum of the dihedral angle metric $D$ over all dihedrals of two conformations) to extend our earlier clustering and nomenclature of antibody CDRs (North, Lehmann, Dunbrack, J. Mol. Biol. 406, 228-256, 2011). For most CDRs and loop lengths, about 10% of the data ends up as noise. An example is shown in the figure below for L3 length 8. The clusters are cleaner in terms of Ramachandran regions, and the sequence profiles are also more predictive in some cases. We have kept our previous nomenclature, which is widely used (L1-11-1, etc.) and created new cluster names and retired others where needed.
83 - Interactome based drug design based on disease-disease relationships

Jonathan Fine1, Joydeb Majumder1, Travis C. Lantz1, Gaurav Chopra1,2,3,4

1Purdue University, 2Purdue Institute for Drug Discovery, 3Purdue Center for Cancer Research, 4Purdue Institute for Inflammation, Immunology and Infectious Disease

Abstract

We have developed an interactome based drug discovery, design, and repurposing platform that analyzes similarity of compound-proteome interaction signatures to determine functional drug behavior, compared to traditional single target approaches, resulting in common links between diseases (prostate cancer, breast cancer, hypertension) to predict repurposeable drugs. Using our new flexible target and antitarget guided design program (CANDOCK), we designed and synthesized potent (IC$_{50} < 1$nM) anticancer drug leads for castration resistant prostate cancer (CRPC) that are non-toxic on normal human prostate epithelial cells and in mice at high dose. We also tested our designed lead in patient-derived xenograft tumors in castrated immune compromised NSG mice outperforming abiraterone, a known CRPC drug. These non-toxic putative drugs are differential nuclear hormone receptor modulators optimized by using the interaction profile (vs single target) of multiple receptors in the androgen signaling pathway to improve efficacy and reduce toxicity. We conclude that compared to traditional single target drug discovery that is slow and error prone, interactome based drug discovery that considers any disease as a heterogeneous combination of other disease mechanisms will have a broader impact leading to chemical control of biological pathways in diseases with overlapping mechanisms foreshadowing a new era of faster drug discovery and design.

Introduction

Modern drug discovery uses large libraries to screen against one or few disease targets using high-throughput screening with iterative design endeavors costing millions of dollars. This paradigm is optimized to develop drugs against a single target which was thought necessary to minimize side effects. However, it has an intrinsic high failure rate due to efficacy of wrong target selection, resistance mechanisms of singular targeting or toxicity due to unknown off-target effects. Our work indicates that most human ingestible drugs function as a differential interaction with multiple proteins from different druggable protein classes. Here, we introduce computational methods for compound library
development based on proteome-wide interactions. Instead of screening millions of compounds that are randomly synthesized and have intrinsic high failure rate, one uses virtual screening with multiple proteins combined with target and antitarget based lead optimization to make modular libraries focused on signaling pathways of interest.

Methods

We have previously implemented a modeling pipeline that generates an interaction between 3,733 human approved drugs and 48,278 proteins (~1 billion predicted interactions). Predictions are sorted and ranked by structural proteome interaction signature similarity to all other known drug signatures approved for particular indications suggesting disease-disease relationships and a new CRPC repurposed lead prospectively validated in human cancer cell lines. Next, we used CANDOCK to design chemical analogs from our CRPC repurposed lead based on target and antitarget inhibition profile for pathway specificity to not engage in typical blind synthetic library design based on single target optimization (Figure 1). These designs were tested for anticancer efficacy, metastatic potential on human cancer cell lines in vitro and efficacy and toxicity compared to existing CRPC drug in vivo.

Figure 1. Disease-disease relationship results in repurposed lead. Synthesis of multitarget (target and antitarget) profile based lead optimization results in pathway specific compound library that is potent, non-toxic and disease-specific compared to single target drugs.

Results & Conclusions

We implemented a proteome design method, CANDOCK, to synthesize potent, non-toxic library of differential-targeting (not promiscuous) hormone specific signaling anticancer agents for CRPC. We experimentally validated our selection for enhanced efficacy and toxicity in vivo.
1 - Inhibitors binding to ATPase domain of Glucose-regulated Protein 78 affects substrate-binding domain movement: Molecular dynamics studies

Dr Seema Mishra, Sagara Nuwan Gurusinghe, Rituparna Bhattacharjee

Abstract and Introduction: Glucose-regulated protein 78 (GRP78) is a key endoplasmic reticulum chaperone protein involved in Unfolded Protein Response. It is a universal potential therapeutic protein target for prevention/cure of cancers such as Glioblastoma as well as bacterial and viral infections such as Ebola, influenza and hepatitis. Two major domains in its structure enable it to perform its function. ATP binding to the ATPase domain of GRP78 facilitates the substrate-binding domain (SBD) function. Thereafter, the binding of the exposed hydrophobic residues of the misfolded or unfolded protein to the substrate binding domain of GRP78 enables the substrate protein to fold properly or be degraded. In its allosteric ATPase cycle when ATP is bound to the ATPase domain, the SBDα subdomain lid is open; in this conformation, SBD has low affinity to substrate. Upon ATP hydrolysis, ADP is bound to the ATPase domain and the lid closes on the bound substrate, leading to a high affinity substrate-binding conformation. Again, after exchange of ADP for ATP, SBDα lid opens, substrate is released which is free to fold then. ATP-competitive inhibitors (−)-epigallocatechin gallate (EGCG) and OSU-03012 have been shown to bind to ATPase domain of GRP78 in vitro (1, 2) and in silico and inhibit its activity. Amino acid residues unique to GRP78 are Ile61, Glu293, Arg297, and Arg367 which are different from HSP70-1A highly conserved ATPAse active site. These residues played a major role in the intermolecular interactions with these inhibitors imparting selectivity to GRP78 (3). We further wanted to assess how EGCG binding to ATPase domain affects substrate binding domain movements. We also wanted to interrogate mechanisms of EGCG action on a structural basis whether it is similar to or different from the allosteric ATPase cycle.
Results and Conclusions

Given that a full-length structure of human GRP78 is now available (PDB ID 5E84), molecular dynamics studies were carried out to assess the domain movements and conformational changes in the overall protein. 50 ns MD simulations with Gromacs 4.6.3 were carried out using free protein, protein bound to ATP and docked to EGCG. Comparison studies in all these cases will be presented. RMSD plot (Figure 1), RMSF plot, number of hydrogen bonds and radius of gyration plots will be generated and discussed. Structural changes in the corresponding domains, more specifically the substrate-binding domain will be highlighted and analysed. Free energy or binding energy calculations using MM-PBSA will be discussed. Key research findings from these molecular dynamics studies will be disseminated to the scientific community.

Figure 1: RMSD vs. time plot of the backbone atoms of GRP78 alone (black), GRP78-ATP complex (red) (work to be continued), and GRP78-EGCG complex (green) relative to energy minimized crystal structure and docked structures for 50 ns simulation

Keywords

Molecular Dynamics
Gromacs

Glucose-regulated protein 78 (GRP78)

Glioblastoma

(−)-epigallocatechin gallate (EGCG)

ATPase domain

Substrate-binding domain

References


2 - Uncovering complex structural relationships of proteins

Aleksandar Poleksic¹

¹University of Northern Iowa

Abstract

Finding an optimal, pairwise alignment of protein tertiary structures is challenging but important task in biomedical research. A wide array of structural changes can make the structural relationship between the input proteins undetectable by even the most accurate methods for protein structure comparison. We present a computational method capable of recognizing complex protein structure relationships of different types. By combining alignment relaxation techniques with multiple structural superpositions and by utilizing a unique dynamic programming technique, our method is able to compute biologically meaningful alignments of proteins that have undergone large conformational changes, domain swaps, circular and scrambled permutations.

Methods

To detect the exact nature of the structural relationship between the input proteins, our method incorporates three different techniques: 1) recursive alignment relaxation and refinement, 2) multiple spatial superpositions, and 3) a unique dynamic programming algorithm for finding circular and scrambled permutations. The technique that yields the overall best (normalized) alignment score is selected to calculate the final alignment¹.

The recursive alignment relaxation yields the highest scores when applied to pairs of structures that exhibit mild to moderate conformational variations. In this technique, the proteins are recursively cut into fragments which are then realigned and pieced together to form a global alignment of input structures. The recursion terminates when the score of the alignment between the uncut, rigid structures exceeds the score of the corresponding flexible alignment (composed by piecing together the two alignment segments).
A different approach is applied to align structures that exhibit large conformational variations (e.g. to align structures of the same or similar proteins in ligand bound and ligand free form). Multiple superpositions are first generated to match all locally similar regions. This process yields multiple distance matrices, one for each pair of locally aligned subsegments. The individual, smaller matrices are concatenated into a single global matrix, which is then given as the input to the dynamic programming routine to find the final alignment.

To detect similarities of structures related by circular or scrambled permutations, the algorithm repeatedly perturbs the dynamic programming matrix by moving the front column to the end. The highest scoring alignment, among all alignments generated in this manner, is selected to represent the best alignment between circularly permuted structures (Fig. 1).

![Figure 1. Perturbing the dynamic programming matrix on the fly to find an optimal alignment of the RIPC structures d1qasa2 (a1-a2) and d1rsya_ (b1-b2).](image)

**Results**

Our method consistently demonstrates the highest accuracy in some widely known benchmarks for protein structure comparison. For instance, our alignments perfectly match the manually curated RIPC alignments\(^2\) in 13 out of 22 test cases. In contrast, the
closest competing algorithm finds 8 such alignments\textsuperscript{1}. In the Sisyphus benchmark\textsuperscript{3}, our alignments perfectly agree with the reference alignments in 37 out of 98 cases, whereas the next best algorithm matches 24 out of 98 pairs\textsuperscript{1}. The overall accuracy of our method, when computed on all tested pairs in the two benchmarks, follows a similar trend.

References


3 - Network approach integrates 3D structural and sequence data to improve protein structural comparison

Fazle E. Faisal1, Julie L. Chaney1,2, Khalique Newaz1, Jun Li1, Scott J. Emrich1, Patricia L. Clark1, Tijana Milenkovic1

1University of Notre Dame, 2Siemens Healthineers, Elkhart, Indiana

Proteins are macromolecules that keep us alive. Understanding protein function experimentally is expensive and time-consuming. Consequently, computational prediction of protein function has received attention. In this context, protein structural comparison (PC) aims to quantify similarity between proteins with respect to their structural patterns, in order to predict functions of unannotated proteins based on functions of annotated proteins that they are structurally similar to. Initial PC approaches were based on sequence patterns. Since amino acids that are distant in the sequence can be close in the 3-dimensional (3D) structure, 3D contact approaches can complement sequence approaches. Traditional 3D contact approaches study 3D structures directly. Instead, 3D structures can be modeled as protein structure networks (PSNs). Then, network approaches can compare proteins by comparing their PSNs. Network approaches may improve upon traditional 3D contact approaches. We cannot use existing PSN approaches to test this, because: 1) They rely on naive measures (patterns) of network topology. They cannot integrate 2) multiple PSN measures or 3) PSN data with sequence data, although this could help because the different data types capture complementary biological knowledge. We address this by: 1) exploiting well-established graphlet (aka "network motif") measures via a new network approach, which allows for 2) integrating multiple PSN measures, and 3) combining the complementary PSN data and sequence data. We compare both synthetic networks and real-world PSNs more accurately and faster than existing network, 3D contact, or sequence approaches.

Figure 1: (a) Illustration of how a protein, whose amino acid sequence folds into a 3-dimensional (3D) structure can be modeled as a protein structure network (PSN). In the
PSN, nodes are amino acids and two amino acids are linked by an edge if they are spatially close enough in the 3D structure. (b) Traditional approaches for protein structural comparison (PC) have been sequence- or 3D structure-based, while our approach is PSN-based. While some PSN-based approaches for PC do exist, they are typically based on primitive measures of network topology. On the other hand, our approach is based on graphlets (small subgraphs aka “motifs”) and is thus more powerful. (c) Illustration of the accuracy (in terms of the area under precision-recall (AUPR) or ROC (AUC) curve) of different method types in the task of PC, for two representative protein data sets. In each column, the best method is bolded. Similar results hold for all other analyzed data sets.
4 - Drug search for leishmaniasis: a structure-based
drug discovery approach for detecting anti-Leishmania
hits

Rodrigo Ochoa\[^{1,2}\], Stanley J. Watowich\[^{3}\], Andrés Flórez\[^{4}\], Sara M. Robledo\[^{1}\], Carlos Muskus\[^{1}\]

\[^{1}\]Programa de Estudio y Control de Enfermedades Tropicales -PECET, Universidad de
Antioquia, Medellín, Colombia, \[^{2}\]Max Planck Tandem Group, Biophysics of Tropical
Diseases, University of Antioquia, Medellín, Colombia, \[^{3}\]Department of Biochemistry and
Molecular Biology. University of Texas Medical Branch –UTMB, Galveston, United
States, \[^{4}\]Division Theoretical Systems Biology. German Cancer Research Center
(DKFZ), Heidelberg, Germany

**Introduction**

Leishmaniasis is a neglected tropical disease treated mainly by chemotherapy options
[1]. However, it is challenging due to the increase in drug resistance, cost and the lack of
available treatments. In addition, side and adverse effects have been associated,
including death [2].

Thanks to the availability of crystallized proteins solved for different species of the
parasite, we implemented a structure-based drug discovery approach combining the
advantages of molecular dynamics (MD) simulations and molecular docking. The project
aimed to identify hit candidates in collaboration with the World Community Grid (WCG),
using an open platform to distribute the computational simulation through integrated
machines from different parts of the world.

**Methods**

The structures of 53 *Leishmania spp.* proteins were selected and downloaded from the
Protein Data Bank (PDB) database [3]. Each protein was carefully reviewed in order to
maintain required external molecules and biological conditions necessary for more
accurate simulations. Initially, the proteins were submitted to MD simulations of five
nanoseconds using GROMACS v5.0 [4], with the aim to recreate the protein flexibility.
From each MD trajectory, a set of 10 snapshots were docked using the AutoDock Vina
software [5]. The screening was made against a library of approximately 600,000 drug-
like compounds from the ZINC database based on a Relaxed Complex Scheme (RCS) methodology. Finally, five of these compounds targeted against DHODH were evaluated in vitro using different human macrophage cell lines and GFP *Leishmania* parasites.

**Results and conclusions**

Most of the docking hits were detected against three proteins: UDP-glucose pyrophosphorylase, dihydroorotate dehydrogenase (DHODH) and tyrosyl-tRNA synthetase. An inspection of one of the hits (C01) against the DHODH is shown in Figure 1.

**Figure 1. Docking analysis for hit C01 bound to LmDHODH. a) Predicted binding of C01 to the LmDHODH crystal structure. B) Superimposed conformations of C01 predicted to bind to each of the LmDHODH conformations extracted from the MD simulation.**

*In vitro* testing of eight of the best compounds detected by docking against the DHODH, showed that three had acceptable activity against *Leishmania* parasites, and one had a strong activity to kill the parasites without affecting the human-derived cell lines. However, these compounds can be optimized to increase their activity. In addition, new compounds from the *in silico* search can be also evaluated in vitro, and the hits with good selectivity index should be evaluated in vivo.

After finishing the virtual screening post-analysis, a website (http://ubmc-pecet.udea.edu.co/index.php/dsfl) with all the data about the used proteins, their hits and the docking scores is available for the public, in order to accelerate the drug discovery pipeline for neglected tropical diseases such as leishmaniasis.
References


5 - Deep learning based subdivision approach for large scale macromolecules structure recovery from electron cryo tomograms

Min Xu¹

¹Carnegie Mellon University

Motivation

Cellular Electron CryoTomography (CECT) enables 3D visualization of cellular organization at near-native state and in sub-molecular resolution, making it a powerful tool for analyzing structures of macromolecular complexes and their spatial organizations inside single cells. However, high degree of structural complexity together with practical imaging limitations make the systematic de novo discovery of structures within cells challenging. It would likely require averaging and classifying millions of subtomograms potentially containing hundreds of highly heterogeneous structural classes. Although it is no longer difficult to acquire CECT data containing such amount of subtomograms due to advances in data acquisition automation, existing computational approaches have very limited scalability or discrimination ability, making them incapable of processing such amount of data.

Results

To complement existing approaches, in this paper we propose a new approach for subdividing subtomograms into smaller but relatively homogeneous subsets. The structures in these subsets can then be separately recovered using existing computation intensive methods. Our approach is based on supervised structural feature extraction using deep learning, in combination with unsupervised clustering and reference-free classification. Our experiments show that, compared to existing unsupervised rotation invariant feature and pose-normalization based approaches, our new approach achieves significant improvements in both discrimination ability and scalability. More importantly, our new approach is able to discover new structural classes and recover structures that do not exist in training data.

Availability

Source code freely available at http://www.cs.cmu.edu/~mxu1/software
6 - Use of cross-docking simulations for identification of protein-protein interactions sites: the case of proteins with multiple binding sites

Nathalie Lagarde1, Lydie Vamparys1, Benoist Laurent1, Alessandra Carbone2, Sophie Sacquin-Mora1

1Laboratoire de Biochimie Théorique, CNRS UPR9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005, Paris, France, 2Laboratoire de Biologie Computationnelle et Quantitative, CNRS UMR7238, UPMC Univ-Paris 6, Sorbonne Université, 15 rue de l'Ecole de Médecine, 75006, Paris, France

Abstract

Protein–protein interactions (PPI) play a central role in biological systems. To evaluate the ability of a cross-docking method to detect multiple binding sites on protein surfaces, we used the MAXDo algorithm with a rigid-body docking approach and a coarse-grain protein model. Alternative interfaces different from the reference experimental interfaces were predicted and turned out to be interfaces with other partners. We compared the use of two different scoring schemes accounting for multiple binding sites, for evaluating the binding sites prediction.

Introduction

Understanding PPI is essential to decipher the mechanism of numerous biological functions. Some in silico methods can be used to investigate PPI. In particular, cross-docking simulations of large datasets of proteins can be used to predict interface residues1-3. Many proteins present multiple binding sites on their surfaces and are particularly challenging for PPI studies. We discuss here the ability of the cross-docking method to detect multiple binding sites on protein surfaces.

Methods

358 proteins extracted from 138 unique PDB structures were used for this study. The MAXDo algorithm1 was used with a rigid-body docking approach and a reduced protein representation, a coarse-grain protein model developed by Zacharias4. Binding site predictions resulting from evolutionary sequence analysis produced with JET5 were used.
to restrict the initial search space. For each surface residue, its Protein Interface Propensity (PIP) was computed and used to predict binding sites on the protein surface.

Results & Conclusions

For a large number of proteins, alternative interfaces different from the reference experimental interfaces were predicted. However, about 70% of these interfaces were not false positives but correspond to interfaces with alternate partners (Figure 1).

**Figure 1.** Mapping the PIP values on a protein’s surface, high PIP residues are shown in blue and low PIP residues are shown in red. The reference experimental partner (ref) is shown in a black cartoon representation, while the alternate partner (alt) is shown in green. (A) Interface with an other chain of the same PDB structure not included in the database. Vimentin (1GK4_F) in complex with other vimentin monomers (1GK4_A (ref) and 1GK4_E (alt)); (B) Interface with a nucleic acid molecule. Transcription factor SOX-2 (1GT0_D) in complex with POU domain (1GT0_C, ref) and DNA (alt); (C) Homodimerization interface. HIV-1 gp41 glycoprotein (1AIK_N) in complex with two others HIV-1 gp41 glycoprotein monomers: 1AIK_C (ref) and 1AIK_D (alt) identified using the PIQSI database.

We compared the use of two different scoring schemes accounting for multiple binding sites, for evaluating the binding sites prediction. The first score was obtained by comparing the predicted interface with one single global reference experimental interface generated by concatenating all the existing experimental interfaces. In the second score, the predicted interface was compared to each experimental interface separately, and only the interface associated with the best performance was kept.
Using cross-docking simulations on a large dataset of proteins, accurate binding site predictions could be realized, including proteins which present multiple binding sites.

References


7 - Using Ancestral Sequence Reconstruction to Characterize an Allosteric Bi-Enzyme Complex

Kristina Heyn¹, Alexandra Holinski¹, Rainer Merkl¹, Reinhard Sterner¹

¹University of Regensburg

Ancestral sequence reconstruction (ASR) is the inference of primordial amino acid sequences from contemporary ones with the help of a phylogenetic tree [1]. Extant sequences occupy the leaves of this tree and the sequences corresponding to the internal nodes and the root are intermediates. The in silico and biochemical characterization of these intermediates can help to elucidate the structural determinants of a whole protein family.

Imidazole glycerol phosphate synthase is a bi-enzyme complex consisting of the cyclase subunit HisF and the glutaminase subunit HisH. We have used ASR and protein design to unravel the structural basis of the HisH-HisF interaction between the two enzymes.

To this end, we compared the binding of a given HisH protein to i) a modern HisF enzyme (a leaf), to ii) evolutionary intermediate HisF proteins (internal nodes), and to iii) HisF from the last universal common ancestor (LUCA-HisF, root) that differ in the composition of their interfaces [3].

The in silico analyses of these interfaces made clear that one residue is key to the binding affinity of the HisH-HisF complex. The predicted effect on complex stability induced by the reciprocal exchange of the corresponding interface residues was confirmed by means of biochemical studies [4]. Thus, we could demonstrate that a combination of in silico methods and wet-lab experiments allowed us to identify an interface hot-spot in a straightforward manner.

References


8 - Structure-based drug repositioning identifies novel Hsp27 inhibitors, which efficiently suppress drug resistance development in cancer cells

Michael Schroeder¹, Jörg Heinrich¹, Sainitin Donakonda¹, Joachim Haupt¹, Yixin Zhang¹, Petra Lennig¹, Sebastian Salentin¹, Melissa Adasme¹

¹TU Dresden

Drug resistance is an important open problem in cancer treatment. In recent years, the heat shock protein HSP27 (HSPB1) was identified as a key player driving resistance development. HSP27 is overexpressed in many cancer types and influences cellular processes such as apoptosis, DNA repair, recombination, and formation of metastases. As a result cancer cells are able to suppress apoptosis and develop resistance to cytostatic drugs.

To identify HSP27 inhibitors we follow a novel structure-based drug repositioning approach. We exploit a similarity between a predicted HSP27 binding site to a viral thymidine kinase to generate lead inhibitors for HSP27. We characterise binding of a known inhibitor with interactions patterns of our tool Plip and exploit this knowledge to assess better binders.

Six of these leads were verified experimentally. They bind HSP27 and down-regulate its chaperone activity. Most importantly, all six compounds inhibit development of drug resistance in cellular assays. One of the leads – chlorpromazine – is an antipsychotic, which has a positive effect on survival time in human breast cancer. In summary, we make two important contributions: First, we put forward six novel leads, which inhibit HSP27 and tackle drug resistance. Second, we demonstrate the power of structure-based drug repositioning. The identified compounds will now undergo preclinical studies.

The work was published in

1. OncoTarget (Impact factor 5): Heinrich et al., New HSP27 inhibitors efficiently suppress drug resistance development in cancer cells, OncoTarget, June 2016) and

The talk will give an exiting account of how research in the 80s into a herpes drug led to the recent discovery of novel anti-cancer candidate compounds. The talk will give details on algorithms for structure-based binding site comparison, which is at the heart of the discovery. It will illustrate how the method implements scaffold hopping, which leads to completely new ligands not anticipated before. It will also sketch the limits of the approaches including the role of modelling 3D structures, binding site prediction, and docking.
9 - Zipping and assembly with limited sets of constraints

Maryana Wånggren¹, Martin Billeter², Graham J.L. Kemp¹

¹Computer Science and Engineering, Chalmers University of Technology, ²Department of Chemistry and Molecular Biology University of Gothenburg

Abstract

We aim to improve the speed and accuracy with which three-dimensional models of protein structures can be generated from easily and rapidly obtainable nuclear magnetic resonance (NMR) data. We use a zipping and assembly dynamic programming approach to search for protein conformations that are consistent with given distance and angle constraints. Our approach benefits from having both high level and low level descriptions of conformational features and constraints, and the possibility to infer new constraints from those that are given.

Introduction

NMR experiments provide a variety of restraints that can be used when constructing a model structure. Some experiments are relatively straightforward and are routinely performed when studying a new protein. Residual dipolar couplings (RDCs) are the easiest data to obtain for large proteins and give restraints related to torsion angles and also to the orientation (in a fixed coordinate system) of selected bonds, often the N-H bonds of the backbone. Other experiments require alternative isotope labelling or multidimensional NMR, which make the experiments very time-consuming and more expensive. One is often faced with limited and sometimes insufficient information for determining a well-resolved 3D structure. In addition, the type of data available for different proteins may vary: ranges for torsion angles, distance approximations, relative orientation of different molecular parts etc. We want to build accurate model structures quickly (a few minutes), using only data that are easy to obtain.

Methods

A protein modelling program has been implemented [1] that uses the zipping and assembly method in which longer fragments are constructed from pairs of shorter ones [2]. The built fragments must be free from steric clashes, and be compatible with given constraints (see below). Distance constraints can be propagated: new constraints lower
in the zipping and assembly data structure can guide the conformational search towards feasible solutions.

Results

We are currently testing our method with a range of proteins by generating ensembles of structures based on only secondary structure information and disulphide bridges. The resulting models are compared with experimentally determined structures from the Protein Data Bank. As an illustration, Figure 1(A) shows distance constraints for human beta-defensin 6 [3] mapped onto the zipping and assembly data structure. An ensemble of models that are compatible with these constraints is shown in Figure 1(B).

![Figure 1](image)

**Figure 1.** (A) Distance constraints for human beta-defensin 6 mapped onto cells in the zipping and assembly data structure. Cells labelled “S” (gold) indicate three pairs of residues that form disulphide bonds: (6,33), (13,27), (17,34). Cells labelled “A” (blue) indicate antiparallel bridges identified by HN-HN NOEs: (12,34), (14,32), (22,35), (25,33). Another distance constraint between positions 6 and 17 (green) can be inferred from these. Additional distance constraints between other pairs of residues (unlabelled cyan and yellow cells) can be inferred from the eight constraints listed above. An alpha-helix, derived from chemical shift data, gives distance constraints between residues i and i+4 (magenta). (B) Alpha-carbon traces of 50 models of human beta-defensin 6. The structures of the core are in good agreement and the unconstrained terminal regions are dynamic.

Acknowledgements

This work is supported by a Project Research Grant from VR (621-2011-6171).

References


12 - Recurrent neural network models to quantitatively predict RNA-RNA interactions

Michelle Wu1, Johan Andreasson2,3, Wipapat Kladwang2, William Greenleaf3, Rhiju Das2,4

1Biomedical Informatics Training Program, Stanford University School of Medicine, 2Department of Biochemistry, Stanford University School of Medicine, 3Department of Genetics, Stanford University School of Medicine, 4Department of Physics, Stanford University

Background

RNA is a functionally versatile biomolecule that plays a key role in myriad novel RNA-based synthetic biology tools, including designer riboswitches controlling living cells, CRISPR/Cas9-based genome editing, RNA-based silencing, and reengineering of mRNA translation. The ability to precisely design nucleic acid interactions is critical to gaining precise control over these systems. However, doing so relies upon a thorough understanding of the energetics of RNA interactions, and existing computational models have proven inadequate to quantitatively account for the biophysical behavior of RNA molecules. Recent technological advances have allowed for the collection of quantitative biophysical information on millions of individual RNA targets [1]. Further, algorithmic advances in the field of deep learning have demonstrated unprecedented power in extracting the relevant features from large datasets [2]. We combine these high-throughput biophysical datasets with recurrent neural networks to build quantitative models of RNA interactions.

Results

Tens of thousands of RNA molecules were designed by citizen scientists through the Eterna open laboratory [3]. These sequences were designed to modulate their affinity to an RNA reporter in response to the concentrations of three RNA molecules whose expression levels are indicative of active tuberculosis [4]. The design-reporter affinities were measured in up to 16 different sets of concentrations of the three input RNAs using a massively parallel RNA array platform [1], resulting in over 7 million individual affinity measurements. These $K_d$ values were used to train a recurrent neural network model to predict reporter affinity to any design sequence in the presence of the three input RNAs at arbitrary concentrations. This model was able to achieve blind predictions with an RMSE of 0.88 kcal/mol on a test set that included conditions not seen in the training
data. Further, we show that data augmentation methods specific to sequence data are able to improve the performance to 0.78 kcal/mol.

Conclusions

In this work, we show that recurrent neural networks are able to quantitatively predict the affinities of RNA-RNA interactions. This suggests that these models have the potential to capture complex biophysical information and may enable precise design of RNA interactions for biotechnology applications.

References


13 - SPRINT: Ultrafast protein-protein interaction prediction of the entire human interactome

Yiwei Li\textsuperscript{1}, Lucian Ilie\textsuperscript{1}

\textsuperscript{1}University of Western Ontario

Abstract

We present SPRINT (Scoring PRotein INTeractions), a new sequence-based algorithm and tool for predicting protein-protein interactions (PPI). We comprehensively compare SPRINT with state-of-the-art programs on seven most reliable human PPI datasets and show that it is more accurate while running five orders of magnitude faster.

Introduction

Proteins perform their functions usually by interacting with other proteins. Predicting which proteins interact is a fundamental problem. Experimental methods are slow, expensive, and have a high error rate. Many computational methods have been proposed among which sequence-based ones are very promising. However, so far, no such method can predict effectively the entire human interactome.

Methods

The basic idea of SPRINT is that proteins similar with interacting proteins are likely to interact as well. That is, if proteins $P_1$ and $P_2$ are known to interact and the sequence of $P_1$ ($P_2$) is similar with that of $P_1'$ ($P_2'$), then $P_1'$ and $P_2'$ are likely to interact as well. SPRINT uses a highly efficient, spaced-seed-based algorithm to quickly find all relevant similarities among all protein sequences, then employs an effective scoring scheme to evaluate the contribution of the similarities discovered to the likelihood of interaction.

Results & Conclusions

We compared SPRINT with five state-of-the-art programs for PPI prediction on seven most reliable human PPI datasets available. The competing methods are developed by Martin\textsuperscript{2}, Shen\textsuperscript{3}, Guo\textsuperscript{4}, Pitre\textsuperscript{5} (PIPE2), and Ding\textsuperscript{6}.
We first evaluate the programs on the breakthrough datasets of Park and Marcotte\textsuperscript{1}. For each training set, Park and Marcotte created three testing sets, depending on whether both proteins in each testing pair appear in training (C1), only one appears (C2), or none (C3). This distinction is essential in assessing the ability of PPI programs to generalize well to full population level. The receiver operating characteristic (ROC) and precision-recall (PR) curves for each class are shown in Figure 1.

![Figure 1: Comparison on Park and Marcotte datasets: ROC and PR curves](image)

**FIGURE 1.** Comparison on Park and Marcotte datasets: ROC and PR curves

We then compared the top four programs and included six human datasets: Biogrid, HPRD, InnateDB manually curated, InnateDB experimentally validated, IntAct, and MINT.

For prediction purposes, the most important behaviour is for high specificity. We show in Figure 2, for several high specificity values, the average sensitivity, precision, and F\textsubscript{1}-score. The average is computed over all seven datasets. Overall averages, over the three classes, are also given. SPRINT clearly outperforms the other programs.

The average area under ROC and PR curves are given in Figure 3, including overall averages. SPRINT again outperforms the competition.
In practice, these programs are supposed to predict the entire human interactome, that is, training on an entire dataset, then predict the probability of interaction of ALL protein pairs. SPRINT requires 15 to 100 minutes to finish this task, depending on the initial dataset, while the other programs require months or years.

In conclusion, SPRINT is a more accurate and much faster sequence-based PPI prediction algorithm and tool. Our goal is to transform the very challenging problem of predicting the entire human interactome into a routine task.

References


14 - Seeing the Trees through the Forest: Sequence-based Homo- and Heteromeric Protein-protein Interaction sites prediction using Random Forest

Qingzhen Hou¹, Paul F. G. De Geest¹, Wim F. Vranken², Jaap Heringa¹, K. Anton Feenstra¹

¹VU University, Amsterdam, ²Vrije Universiteit Brussels

Introduction

With ever-increasing numbers of protein sequences becoming available, annotating genome data is a challenging task. Protein-protein interactions (PPIs) play a central role in virtually all cellular processes. Identification of interface (IF) sites between interacting proteins is essential to understand complex formation and investigate their functions. Few of the available genomes, however, have a good coverage of experimentally validated annotations for PPIs. This makes the prediction of protein interaction sites from sequence information increasingly attractive. We here explore the potential of sequence information and derived properties of interacting residues as features to predict interacting amino acids.(1)

Methods

We used Random Forest (RF) classifiers with as features the surface accessibility and secondary structure from NetsurfP,(2) and backbone flexibility (N-H S²) from DynaMine,(3) for the query, and average and standard deviation of features and sequence entropy for homologs from Blast, over a 9-residue window. We trained it on homodimeric, heterodimeric and mixed datasets, and tested it on strictly independent data; these were based on Test set 1 (4) and Dset 186 and Dset 72.(5)

Figure 1: Comparison of performance between our methods and PSIVER on the heterodimer testset. ROC plots and Precision-Recall plots are used to measure
performance of IF site prediction. Stars '*' in the figure indicate the point for each method using their own default thresholds.

Results

Figure 1 shows a comparison of RF\textit{\_homo} and RF\textit{\_hetero}, trained on homomeric and heteromeric IFs, respectively, with the best sequence-only predictor, PSIVER (5). Our RF\textit{\_hetero} predictor performs best on precision, MCC, F1 and AUC-ROC (0.652) on the independent heteromeric test-set (Dset\_48). We additionally trained a general IF predictor, RF\textit{\_combined}, on the combined homomeric and heteromeric datasets. On the heteromeric test-set (Dset\_48) it performs quite well (AUC-ROC 0.636; PSIVER 0.614), while on the independent homodimer test-set (HM\_479 test) it outperforms the other methods on recall, precision, MCC, F1 and AUC-ROC (AUC-ROC 0.724; PSIVER 0.546).

Discussion

The success of our random forest model suggests that homodimer and heterodimer interfaces have much more in common than is often assumed. As the method is sequence-based and not sensitive to the type of interaction, we expect our research to be of interest to many biological and biomedical researchers in academia and industry, especially when protein structure data is unavailable. For non-expert users we implemented the SeRenDIP web-server (www.ibi.vu.nl/programs/serendipwww/), which only needs a single sequence as input to predict for a protein of interest the most likely interaction position.

References


15 - A computational design strategy for discovering the cellular targets of histone lysine methyltransferases

Diego Alonso-Martinez¹,², Peter A. DiMaggio¹

¹Imperial College London, ²MRC London Institute of Medical Sciences (LMS)

Abstract

The activity of histone lysine methyltransferases (HKMTs) is crucial for the regulation of many gene expression programs and the development and progression of many diseases such as cancer. Currently, there are no biologically-consistent methods to fully characterise the specific protein targets of given HKMTs. Here we present a novel HKMT “methylome” profiling assay, which utilises mixed-integer programming to engineer synthetic cofactors selective for a given HKMT based on structural differences, molecular dynamics (MD) simulations to rank cofactor designs and mass spectrometry (MS) proteomics for target identification.

Introduction

Histone post-translational modifications (PTMs) are dynamic epigenetic marks that play a central role in the regulation of gene expression and are catalysed by various classes of enzymes, including histone lysine methyltransferases (HKMTs). HKMTs catalyse the transfer of a methyl group from a S-adenosyl methionine (SAM) cofactor to a specific histone lysine target residue. Since all HKMTs utilise the same SAM cofactor and many can modify a given target residue, there is currently no way to directly assess which HKMT performs an observed methylation event. This technology gap limits our ability to develop more specific treatments for epigenetically-regulated diseases, such as cancer. Here we propose to engineer a cellular assay to selectively profile the methylation targets of sets of HKMTs.

Methods

Our proposed strategy is to identify existing structural differences between the catalytic SET domains of different HKMTs that could selectively accommodate synthetic SAM cofactors. Specifically, the approach aims to identify unique “holes” in HKMT structures and then design synthetic SAM cofactors with a complementary “bump” so that it binds
selectively with that HKMT. The corresponding bumped SAM analogues also contain a heavy methyl donor group (e.g. SAM-CD₃) so that a methylation event is biologically consistent and detectable by MS proteomics. The computational design approach takes aligned HKMT-cofactor structures as input and utilises mixed-integer programming to identify pockets that are unique to one HKMT. The output of the program elucidates design points for incorporation of a synthetic “bump” on the SAM analogue, and the candidate analogues are computationally screened for binding affinity and conformational suitability using molecular dynamics. The highest-ranking SAM analogues are then synthesised and used in an in vitro assay to demonstrate selective utilisation by the HKMT of interest,
where mass spectrometry (MS) proteomics is the primary readout for validating catalytic activity. The overall methodology is summarised in Figure 1.

Results & Conclusions

The program identified a natural pocket in the SET domains of HKMTs lacking a post-SET domain (ie SETD7) which does not exist in post-SET containing HKMTs. Several SAM analogues were proposed based on the steric and electrostatic properties of the SETD7 binding pocket and synthesized for experimental validation (in collaboration with Dr Matthew J. Fuchter and Katie Addison, Imperial College London). The proposed SAM cofactors were docked in both SETD7 and G9a (a post-SET containing HKMT) in Autodock Vina 2, followed by molecular dynamics simulations using Amber16 3 and computational assessment of the binding between the cofactor analogues with SETD7 compared to G9a. This computational screening revealed that the binding affinities and orientation of the SAM analogues was similar to that of native SAM for SETD7, but not for G9a. Subsequent in vitro experiments utilising recombinant G9a and SETD7, our engineered SAM analogues and chromatin substrates further validated cofactor selectivity for SETD7, which supports the feasibility of our proposed design methodology.

References

16 - Engineering Improvement of a Potent Human-Derived Monoclonal Antibody Against Respiratory Syncytial Virus Using Structure-Based Computer Modeling

Sean Le¹,², Andreas Loos²

¹Thomas Jefferson High School for Science and Tech, ²Aridis Pharmaceuticals

Respiratory syncytial virus (RSV) infections are a common cause of hospitalizations in infants, resulting in 57,000-120,000 hospitalizations in the U.S. annually. Synagis®, the current therapy for RSV infection prevention has limited effectiveness, and must be given five-six times during the RSV season. A recently discovered monoclonal antibody, AR-201, has shown to be more potent than Synagis®. Our goal is to use structure-based computer modeling to engineer AR-201 to prolong its time (half-life) in the blood, thus reducing its dosing frequency to optimally once during the RSV season. Our approach to improving AR-201 half-life is to modify binding of AR-201 to the neonatal Fc receptor (FcRn), which salvages antibodies from degradation and recycles them to the blood stream. We selected seven published IgG mutants which showed increased half-life in other antibodies. We introduced these mutations into AR-201 and predicted their 3D structures. The quality of the 3D structures was assessed and optimized in order to create more realistic structures. The optimized 3D structures were then docked onto FcRn and their computer models were generated. We calculated the predicted binding affinities between the AR-201 mutants and FcRn. Our results showed mutant C6A-78A to have the highest calculated binding affinity to FcRn, indicating that it has the highest potential for prolonged half-life. This work will guide the further development of AR-201 into a highly potent drug with increased half-life. Although still in its infancy, molecular computer modeling can save substantial time and effort during the drug development process.
17 - An efficient algorithm for improving structure-based prediction of transcription factor binding sites

Jun-tao Guo¹, Alvin Farrel¹

¹University of North Carolina at Charlotte

Abstract

We present here a novel pentamer algorithm for structure-based transcription factor (TF) binding site prediction. The DNA binding sequence is split into overlapping pentamers for calculating TF-pentamer interaction energy, which was then combined for full-length binding site prediction. Our results show that the new pentamer approach improves TF binding site prediction accuracy while dramatically reducing the time complexity, especially for longer binding sites. To our knowledge, this is the first fragment-based method for structure-based TF binding sites prediction.

Introduction

Structure-based TF binding site (TFBS) prediction considers the physical interactions between a TF and candidate binding sequences. The binding affinities between TFs and their binding sequences can be calculated with knowledge-based potentials, physics-based energy or a combination of both [1]. To take advantage of the pros of both knowledge-based and physics based potentials, we developed an integrative energy (IE) function that consists of a residue-level knowledge-based multi-body (MB) potential, an hydrogen bond energy, and a pi-interaction energy, for structure-based prediction of TFBSs [1]. The IE function improved the TFBS prediction accuracy. However, as the size of the binding sites increases, the time complexity increases exponentially.

Here we propose a novel algorithm, by splitting the DNA binding sequence into pentamers, for more efficient and accurate TFBS prediction. To our knowledge, this is the first attempt to use DNA binding fragments for structure-based TF binding site prediction. Results show that our new approach not only dramatically improve the prediction speed, it also helps improve prediction accuracy, especially for TF dimers with longer binding sites.
Methods

The DNA sequence is split into overlapping pentamers and each pentamer is mutated to every possible permutation. The binding energy for each TF-pentamer DNA complex is calculated using the IE function. The full-length TF binding motif was predicted from these TF-pentamer interaction energies using either tiling array or PWM stacking algorithms. The new method was tested on a non-redundant set of 27 TF monomer-DNA and 8 TF dimer-DNA complexes. IC-weighted PCC, a PWM comparison method to measure the similarity of the corresponding columns between the predicted and the reference PWMs, was used to determine the number of correctly predicted positions.

Results & conclusion

We tested the pentamer method on two datasets and compared the new algorithm with our previous full-length method using three energy functions, IE, BM, and DDNA3. Results show that the pentamer method is much faster than the full-length methods and the improvement is even more significant for longer binding sites (Table 1). More importantly, the new method also improved the prediction accuracy significantly. Six of the 8 dimer cases showed much improved TFBS predictions when compared to the full-length prediction method (Fig. 1). To conclude, we developed a fragment-based pentamer algorithm that greatly speeds up TFBS prediction and improves TFBS prediction accuracy. Our results also show that the longer the binding sequences, the more speedup and accuracy can be achieved.

Table 1. Running time (CPU hours) comparison between the pentamer algorithm and the full-length (FL) algorithm with energy functions IE, MB, and DDNA3.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Length (bp)</th>
<th>Pentamer IE</th>
<th>FL IE</th>
<th>FL MB</th>
<th>FL DDNA3</th>
</tr>
</thead>
<tbody>
<tr>
<td>4IQR</td>
<td>8</td>
<td>2.6</td>
<td>206</td>
<td>48</td>
<td>15</td>
</tr>
<tr>
<td>1NLW</td>
<td>8</td>
<td>7.5</td>
<td>166</td>
<td>56</td>
<td>17</td>
</tr>
<tr>
<td>1GU4</td>
<td>12</td>
<td>12.5</td>
<td>10351</td>
<td>3381</td>
<td>1137</td>
</tr>
<tr>
<td>1OZJ</td>
<td>12</td>
<td>18</td>
<td>26126</td>
<td>4301</td>
<td>1408</td>
</tr>
</tbody>
</table>
**Figure 1.** Comparison of the number of correctly predicted columns (A) and TF-binding motifs (B).

**References**

19 - Structural Determination of Chitinase obtained from Bacteria from the African Catfish (Clarias gariepinus)

Ajayi A.A.¹, Onibokun E. A.¹, Adebiyi M.O.¹, Odia T.¹, Olasehinde G.I.¹

¹Covenant University, Ota, Ogun State, Nigeria

Background

Chitinases are proteins that have been applied in agriculture; as a biopesticide for control of plant fungi infections, in medicine; as indicators of fungi infection, and in waste management; for biodegradation of fish waste. The African catfish (Clarias gariepinus) which plays host to chitinolytic bacteria is very readily available and easy to cultivate thus providing a cheap means for chitinase production in commercial quantity.

Materials and Methods

Chitinase obtained from chitinolytic bacteria inhabiting the gut and skin of Catfish were partially purified with ammonium sulphate precipitation. The DNA isolation was carried out according to the protocol described in Trans Easy pure Genomic DNA kit purchased from Trans gene biotech (China) and PCR carried out. The PCR product was sent to GATC Biotech, Germany for sequencing. The primary sequence of the chitinase was determined in silico by homology modelling using the Swiss Prots prediction engine. The alignment was done on the National Centre for Biotechnology Information (NCBI) and the structure of the chitinase obtained from Bacillus circulans, a bacterium found in the soil was used. The organism has been implicated in human infections such as Septicaemia and wound infections. The alignment on NCBI had 98% percent similarity with the sequence query. The Ramachadran plot was used to determine the chitinase structure with the highest quality.

Results

Molecular identification of the chitinase-producing bacteria revealed that the bacterial samples were Bacillus cereus. The model proposed for the chitinase had 96.6 (456/472) of all residues in favoured (98%) regions with 100 % (472/472) of all residues in allowed (>99.8%) regions.
Conclusion

This study established that species of Bacillus inhabiting the gut and skin of the African catfish can produce Chitinases in appreciable quantity. This study has diversified the use of the African catfish for enzyme production rather than for consumption only. The Structure will help in determining functions associated with the protein.

Keywords

Chitinase, Bacillus sp., kinetic characterization, purification, 3D structure, homology modelling
20 - Drug-target interactoin similairity for drug-target interaction prediction and drug repositioning

Daniele Parisi¹²³, Bart Vanderhoydonck¹, Gert Vriend⁴⁵, Yves Moreau¹²³

¹KU Leuven, ²ESAT, ³STADIUS, ⁴CMBI, ⁵Radboud Universiteit

Abstract

The drug discovery process is long, complex[1] and scarcely productive[2], mostly because of lack of efficacy of the candidate drug[3]. With drug repositioning[4] the information from previous studies could lead a molecule to the market saving half of the time and money. Although the attitude on drug repositioning is optimistic[5] due to some successful projects (i.e. Sildenafil, Thalidomide) and many tools already developed, at the moment this approach does not represent yet a profitable business model[1]. With this work I present some unpublished results to show the potentiality of the collaboration between bioinformaticians and medicinal chemists for a successful drug-repositioning protocol.

Introduction

The computational analysis brought in by Bioinformatics allows to manage massive amount of data, in order to fill sparse matrices about compounds-target interaction[6]. The drug-target interaction predictions are inferred using different techniques (ligand-based, target-based), with specific pros and cons. They have been recently tackled with new chemogenomic methods, called Proteochemometric techniques (PCM)[7], by integrating both the information related to ligands and targets[6]. An example is the ligand-protein interaction profile, a unique and highly informative characteristic that can be easily analysed and compared[8]. In this work I studied and compared drug-target interaction profiles generated with a tool made for bioisosteric replacement, KRIPO[9], to understand their predictive power in drug repositioning. Furthermore I used those predictions in a real case to support a process of hit discovery for new immunosuppressant drugs.

Methods

This work aims to use drug-target interaction similarity for interaction prediction and drug repositioning. Starting from 68.000 PDB complexes, the interaction profiles have been
calculated on the pharmacophores of the interactions between the ligand and the pocket, and stored as fingerprints[9]. Then, each fingerprint has been compared with all the others using a modified Tanimoto Coefficient[9]. The assumption is that pairs ligand-protein with similar interaction profiles would be able to cross-interact, representing a starting point for a repositioning protocol. To evaluate the predictive power, the new predicted couples ligand-protein have been plotted with both their interaction similarity from KRIPO and their experimental activity from CHEMBL. Furthermore this predictive approach has been used to suggest new hit compounds able to inhibit specific kinases involved in the immuno response by the B-cell activation. About hundred compounds with similar interactions have been provided to the medicinal chemists who selected 20 molecules to test against 5 kinases, measuring the level of inhibition.

**Results & Conclusions**

The large scale analysis of predictions showed the weak points of KRIPO in drug-target interaction prediction, due to the excessively long and descriptive fingerprints which reduce the range of similarity to 3 units out of 10(fig1). However, the in vitro tests showed that half of the predicted compounds had a very high level of inhibition ,>75%(fig2). In conclusion, the approach showed here gave interesting results when coupled with experience of chemists but more work is needed to improve the quality of the prediction.

*FIGURE 1*


4. http://doi.org/10.1016/j.drudis.2015.05.001


21 - Frequent Subgraph Mining for Biologically Meaningful Structural Motifs

Sebastian Keller\textsuperscript{1,2}, Pauli Miettinen\textsuperscript{1}, Olga Kalinina\textsuperscript{1}

\textsuperscript{1}Max Planck Institute for Informatics, \textsuperscript{2}Saarland University

Abstract
We present a graph based approach to determine common structural motifs in related proteins using frequent subgraph mining (FSM). To this end, we adapted an existing FSM algorithm to increase its specificity towards biologically relevant and structurally conserved motifs and to make it more lenient towards inaccuracies in biological data.

Introduction
Identification of biologically relevant motifs in protein three-dimensional structures is a long-standing problem in bioinformatics. Here we describe an approach based on FSM, specifically on the gSpan algorithm\cite{1}, to detect such motifs in a given set of related protein structures. The structures are represented as residue interaction networks (RINs), where vertices correspond to residues and edges to interactions between them. FSM then detects subgraphs that have a support above a certain threshold, i.e. are subgraph isomorphic to at least a pre-defined number of RINs. These subgraphs correspond to structural patterns that are strongly conserved among the proteins despite otherwise low sequence or structural similarity, and hence are likely involved in a biological function. Limitations stemming from protein dynamics or experimental procedures of structure determination can result in incomplete RINs in which some subgraphs are not supported due to non-biological reasons.

Methods & Materials
For the first step of converting the structures into RINs, we use RI\textsc{nerator}\cite{2} and label the detected edges as interaction based edges. Additionally, we add sequence based edges between consecutive residues and label them as such. Further we add a second label to each edge corresponding to the euclidean distance between the C-alpha atoms of the residues in order to encode more structural information in the graphs.

To address the issue of the unspecifically reduced support, we introduce the concept of approximately supported subgraphs. It is based on the assumption that artifacts of the
data cause much fewer items to be missing from a RIN across different structures than the lack of a biological feature. Therefore the key criterion for a subgraph to be considered approximately supported is the comparison of its support to the support of all of its parent subgraphs, i.e. all connected subgraphs of the subgraph in question with one edge and perhaps one adjacent vertex removed. We require that the difference in support between the subgraph and each of its parents is less than a low pre-defined threshold.

Utilizing distance-based edge labels without binning also requires changes to the algorithm, because exactly matching distances are not to be expected. To this end, we introduce approximate isomorphism, which allows distance labels of the edges to slightly deviate and still be considered equivalent in the context of isomorphism checks.

We applied our approach to structures from four different protein families: extended AAA-ATPase domain (SCOP: c.37.1.20), eukaryotic proteases (SCOP: b.47.1.2), viral RNA-dependent RNA-polymerase[3] and viral capsids with a jelly roll fold[3].

Results & Conclusion
We rediscover known functional motifs in the first three families and identify a previously undescribed motif in a small evolutionary related subset of capsid proteins. The inclusion of distance-based labels in combination with approximately supported subgraphs allows reducing the number of generic motifs found by the algorithm. Such generic motifs consist mainly of hyrdophobic residues with high propensity to form helices and are therefore not specific to a single protein family.

References
22 - DeepSF: deep convolutional neural network for mapping protein sequences to folds

Jie Hou¹, Badri Adhikari¹, Jianlin Cheng¹

¹Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA

Abstract

Protein fold recognition is an important problem in structural bioinformatics. Almost all traditional fold recognition methods use sequence comparison (alignment) to indirectly predict the fold of a target protein based on the fold of a homologous template protein with known structure, which cannot explain the relationship between sequence and fold. Only a few methods had been developed to classify protein sequences into a small number of folds due to methodological limitations, which are not generally useful in practice. Here, we develop a deep 1D-convolution neural network (DeepSF) to directly classify any protein sequence of arbitrary length into one of 1195 known folds, which is useful for both fold recognition and the study of sequence-structure relationship. Different from traditional sequence alignment based methods, our method automatically extracts fold-related features from a protein sequence of any length and maps it to the fold space with good accuracy. The hidden features extracted from sequences by our method are robust against sequence mutation, insertion, deletion and truncation, and can be used for other protein pattern recognition problems such as protein clustering, comparison and ranking.


Method
Figure 1. The architecture of 1D deep convolutional neural network for fold classification. The network accepts the features of proteins of variable sequence length (L) as input, which are transformed into hidden features by 10 hidden layers of convolutions. Each convolution layer applies 10 filters to the windows of previous layers to generate L hidden features. Two window sizes (6 and 10) are used. The 30 maximum values of hidden values of each filter of the 10th convolution layer are selected by max pooling, which are joined together into one vector by flattening. The hidden features in this vector are fully connected to a hidden layer of 500 nodes, which are fully connected to 1195 output nodes to predict the probability of each of 1195 folds. The output node uses softmax function as activation function, whereas all the nodes in the other layers use rectified linear function \( \max(x, 0) \) as activation function. The features in the convolution layers are normalized by batches.

The architecture of the deep convolutional neural network for mapping protein sequences to folds (DeepSF) is shown in Figure 1. It contains 15 layers including an input layer, 10 convolutional layers, one K-max pooling layer, one flattening layer, one fully-connected hidden layer and an output layer. The input layer has \( L \times 45 \) input numbers representing the positional information of a protein sequence of variable length \( L \). The softmax function is applied to the nodes in the output layer to predict the probability of 1,195 folds.

Results

We train and test our method on the datasets curated from SCOP1.75, yielding a classification accuracy of 80.4\%. On the independent testing dataset curated from SCOP2.06, the classification accuracy is 77.0\%. We compare our method with a top profile-profile alignment method - HHSearch on hard template-based and template-free modeling targets of CASP9-12 in terms of fold recognition accuracy. The accuracy of our method is 14.5\%-29.1\% higher than HHSearch on template-free modeling targets and 4.5\%-16.7\% higher on hard template-based modeling targets for top 1, 5, and 10 predicted folds.
24 - Understanding Protein Interactions at a Molecular Level with Web Tools Using Inter-Residue Contacts and Intermolecular Contact Maps

Romina Oliva¹, Ida Autiero², Zhen Cao², Luigi Cavallo²

¹University Parthenope of Naples, ²KAUST Catalysis Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract

Web tools for the analysis of 3D structures of protein-protein complexes and for the scoring of docking poses are presented, which are based on a novel approach, using inter-residue contacts and their visualization in intermolecular contact maps.

Introduction

Characterization of the interaction interface of protein-protein complexes is a fundamental step for understanding protein interactions at a molecular level. In the many cases where experimental structures are not available, protein-protein docking becomes the method of choice for predicting the arrangement of a complex. However, reliably scoring protein-protein docking poses remains an unsolved problem and the screening of many docking models is thus usually required in the analysis step. In this context, we introduced novel tools for the analysis of structures of protein-protein complexes and for the scoring of docking poses. They rely on the use of inter-residue contacts and their visualization in inter-molecular contact maps and can straightforwardly be applied to protein-nucleic acid complexes.

Methods

We use the conservation of inter-residue contacts at the interface as a measure of the similarity between different protein complex conformations. Contact conservation is then visualized in “consensus contact maps”, i.e. intermolecular contact maps where the more conserved the contact the darker the corresponding dot. CONSRANK first calculates the conservation of each inter-residue contact in the ensemble they belong to, then ranks models based on their ability to match the most conserved contacts.
Results & Conclusions

The first tool we introduced in the field, COCOMAPS (bioCOMplexes COntact MAPS), is a web server to analyse and visualize the interface in biomolecular complexes. COCOMAPS combines the traditional analyses and 3D visualization of the complexes with the effectiveness of the contact map view. It can be applied to the analysis of both experimental and predicted 3D structures.

More recently, we have used inter-residue contacts as the basis for other tools conceived for the analysis of conformational ensembles of protein complexes, such as docking models, NMR conformers and MD snapshots. In particular, CONSRANK (CONSensus RANKing), also available as a web server, is devoted to the scoring of docking models. CONSRANK output includes an interactive 3D representation of the consensus map, a ‘3D consensus map’, where the third dimension is given by the conservation rate of each inter-residue contact (Figure 1).

FIGURE 1. CONSRANK 3D map for the CAPRI target T46, showing the consensus contacts for the ensemble of 387 predictor models (gray), contacts present in the model ranked first (red), and contacts present in the model ranked 100th (blue).
Blind testing in the latest CAPRI (Critical Assessment of PRedicted Interactions) rounds showed CONSRANK to perform competitively with the state-of-the-art energy- or knowledge-based scoring algorithms. A recently modified version of CONSRANK includes a contact-based clustering of the models as a preliminary step of the scoring process\textsuperscript{5}. A dedicated CONSRANK module (consrank-nmr) can be applied to: i) analyse the interface in whole ensembles of NMR complexes, and ii) select a single conformer as the best representative of the overall interface\textsuperscript{3}. Finally, MDcons (Molecular Dynamics CONSensus)\textsuperscript{3} is conceived for the analysis of MD trajectories.

Details on the above tools and results of their application to selected case studies will be presented.

References


26 - A covalent docking approach to simulate the interactions of a novel cephalosporin derivative with beta-lactamases from different sources

Anna Verdino¹, Felicia Zollo¹, Annunziata Soriente¹, Margherita De Rosa¹, Anna Marabotti¹

¹Dept. Chemistry and Biology A. Zambelli, University of Salerno (Italy)

Abstract

We present a computational study made with an innovative approach of covalent docking to simulate the binding of a novel cephalosporin derivative to several beta-lactamases from different organisms and to study their interactions in order to infer information to improve the resistance towards these proteins.

Introduction

Cephalosporins are among the oldest antibiotics discovered. They are still widely used in clinics, but during the years a growing problem has seriously impaired their use: the development of many mechanisms of resistance. For this reason, researchers worldwide are focused to create new cephalosporins able overcome this problem.

We have developed a new cephalosporin in which an additional 2-azetidinone ring with substituents has been linked to the 7-aminocephalosporanic moiety by means of an amide bond. This molecule has shown its efficacy towards Gram+ bacteria including S. aureus, with no cytotoxicity, therefore it has been considered a good starting point to develop an innovative class of bifunctional antibiotics.

The main mechanism by which bacteria can escape the action of cephalosporins is the production of beta-lactamases, enzymes able to hydrolyze the beta-lactam ring, thus deactivating these antibiotics. In fact, the intact beta-lactam ring is essential to perform their function. Therefore, in order to develop new antibiotics able to overcome the microbial resistance, it is essential to understand how they interact with the beta-lactamases. To apply a rational approach for the design of new derivatives of this new class of cephalosporins, we have performed a computational study by simulating the
binding of our new molecule to several beta-lactamases of different species, using a particular approach, covalent docking, able to take into account the covalent bond formed between the antibiotic and the enzyme.

Methods

The structures of several beta-lactamases from Gram+ and Gram- bacteria were selected on the basis of a careful evaluation of their structural quality. The structure of the new cephalosporin molecule was built up by using ChemDraw, with one of the two beta-lactam rings alternately open and the cephalosporanic moiety alternatively charged or neutral. Moreover, the two diastereoisomers obtainable from the chemical synthesis were considered separately. All these structures were used to perform covalent docking against the selected beta-lactamases using the flexible side chain method implemented into a modified version of the popular program AutoDock 4.2. A grid map focused on the active site of each protein was used to set up the calculation. For each complex, 100 docking runs were performed using the AutoDock Lamarckian genetic algorithm. The conformations representative of the best energetic and/or of the most populated cluster of poses were selected and analyzed for their interactions with the enzyme by using Discovery Studio.

Results & Conclusions

Our new cephalosporin derivative binds to all beta-lactamases with a predicted binding energy of about -10 kcal/mol, indicating a potential binding affinity in the low nanomolar range. This energy is similar to the one predicted for known beta-lactam antibiotics. Interestingly, it is possible to note that the predicted binding affinities are generally higher when the conventional beta-lactam ring is bound to the enzyme, whereas the isolated 2-azetidinone ring has a lower binding affinity towards these enzymes. The detailed analysis of the complexes obtained and of the residues involved in different kind of interactions with the chemical groups of the antibiotic suggests which groups might be useful for decreasing the affinity of this compound towards these enzymes, thus developing in the future molecules insensitive to their actions.

Acknowledgements

The present project is funded by FARB (Fondo di Ateneo per la Ricerca di Base), cod. ORSA151138 and ORSA161582 (A.M).
27 - CATH-based protein structure and function analyses to understand the implication of alternative splicing

Su Datt Lam\textsuperscript{1,2}, Jonathan Lees\textsuperscript{1}, Christine Orengo\textsuperscript{1}

\textsuperscript{1}UCL, \textsuperscript{2}National University of Malaysia

Abstract

Alternative splicing (AS) has been suggested as one of the major processes to expand the diversity of proteomes in multicellular organisms. We used domain structure information from CATH to examine the effects of splicing for a set of developmental splice isoforms in human and fly generated by mutually exclusive exon events.

Introduction

Mutually exclusive exons are characterised by coordinated splicing of exons such that only one of the two exons is retained, while the other is spliced out. These events are enriched in proteomics data, suggesting a functional role\textsuperscript{2}. To explore the structural and functional consequences of MXE splicing events in fly and human, we mapped the isoforms to functional families (FunFams) in the CATH domain structure classification. Relatives in FunFams have highly similar structures and functions\textsuperscript{1}.

Methods

Most MXE events do not involve big structural changes of the protein fold, but changes in residue usage. We mapped 1788 fly MXE pairs of isoforms to 66 CATH-FunFams and 63 human MXE pairs to 53 human FunFams (see Fig 1).
to CATH-Gene3D domain FunFams, variable residues are shown in colour.

Splice regions were mapped to known structures where available, or homology models built using the in-house FunMod pipeline. If no model could be built, splice events were aligned to known structures using an in-house HMM-based protocol. We determined whether the variable residues between exon pairs are exposed to solvent and in the vicinity of any known functional sites e.g. catalytic residues (from CSA), protein-protein interaction sites, protein-small molecule interaction sites (from IBIS), and FunSites (in-house predictions, based on conserved sites in FunFam alignments and proven to be enriched with known functional sites).

**Results & Conclusions**

MXE splice regions span from 10 - 200 residues, with a mean value of 38 and <10 variable residues. The FunFams mapped by the splice regions are distributed among 84 CATH superfamilies. Variable residues in the MXE region are more exposed than other mapped residues (p-value <2.2e-16). Although, there are few splice event pairs that lie close to catalytic residues (<4Å), a large proportion of variable residues in fly MXE events lie close to protein-protein interaction, protein-small molecule and FunSites. Similar results were obtained for the human MXE dataset. Furthermore, in fly, more than 79% of variable region changes close to protein interfaces and ligand-binding sites have significant changes in physicochemical properties (measured using the McLachlan
method). In human this is a lower proportion (~65%) but still significant compared to a random model.

The structural and functional information provided with FunFams allows us to explore structure functional changes between isoforms in great detail. For example, the Collapsin Response Mediator Protein is a Dihydropyrimidinase enzyme (EC 3.5.2.2) thought to be involved in centrosome localization, and positive regulation of Notch signalling pathway. We mapped the splice regions to good 3D models built for the FunFam and calculated surface potentials of the two isoforms (using ABPS). The surface potential has major differences between the two isoforms (see Fig 2). There is also a change in the zinc-binding residue 192, (histidine in isoform CRMP-PA and glycine in isoform CRMP-PB), likely to disrupt the interaction of zinc molecules.

**Fig 2** Superposition of splice regions CRMP-PA (orange) and CRMP-PB (green).
Fig 3 Change of histidine in CRMP-PA to glycine and LIGPLOT view of the zinc metal binding.

In summary

Our FunFam pipeline allows us to assess for the first time, structural and functional consequences for the majority of MXE events in whole genomes. Initial results support MXE events as having a number of important roles in cells.

References

29 - Understanding the Molecular Consequences of Genomic Variation Associated with Drug Resistance in Mycobacterium tuberculosis

Stephanie Portelli¹, Jody E. Phelan¹, David B. Ascher², Taane G. Clark¹, Nicholas Furnham¹

¹London School of Hygiene and Tropical Medicine, ²University of Melbourne

Abstract

Resistance to treatments for M. tuberculosis is of serious concern due to the large number of associated deaths in the developing and developed world. Large scale genomic studies have revealed mutations that are associated with drug resistance. The outstanding question is what are the molecular consequences of these variations. Using a range of analyses including in silico predictive tools to estimate the enthalpic effects of point mutations reveals the molecular mechanisms behind mutations that would otherwise be considered inconsequential for introducing therapeutic failure. Ultimately these insights can be used to aid the development of future drugs and, through their integration into predictive tools, in pathogen surveillance.

Introduction

Tuberculosis is a serious airborne infection caused by the bacillus M. tuberculosis and is associated with 1.5 million deaths per year. Of major concern is the increasing drug resistance resulting from patient non-compliance to long-term treatment. Genomic variation associated with underlying resistance to single, multi and extensively drug-resistant tuberculosis was obtained by genome-wide association study of 6,465 clinical isolates from 30 countries. In silico tools were used to estimate the enthalpic effects of point mutations and haplotypes associated with 3 anti-tuberculosis treatments: isoniazid (INH, targets catalase peroxidase); rifampicin (RIF, targets RNA polymerase (RNAP) β-subunit); and D-cycloserine (DCy, targets alanine racemase).

Methods

Molecular models of the functional units of the target proteins, including interactions with binding partners and drug molecules, were built. The latest methods that represent protein residue environments as graphs, where nodes are the atoms and the edges are
the physicochemical interactions established among them and encapsulated as graph-based structural signatures, were used to predict effects brought about via changes in target protein stability and the affinities (as Gibbs free energy) for their drugs, associated proteins, co-factors and DNA. In addition, the effects on function, geometrical distance from interaction regions and experimental phenotypic measures e.g. minimum inhibitory concentrations were integrated.

Results & Discussion

Mutations associated with three therapeutics (INH, RIF and Dcy) that represent first and second line treatments for TB were identified and analysed. For example, 201 mutations associated with RIF drug resistance were identified for the RNAP β-subunit, with 33 residues are associated with multiple SNPs. RNAP β-subunit exists in a complex with α1, α2, β’ and ω subunits, which was modelled in its entirety, in addition to modelling its DNA substrate and its interaction with RIF. The effects of mutations on overall stability as well as the stability of protein-protein/DNA/ligand interactions were calculated and mapped to the structure (Figure 1). The most frequent mutation is S422L, with a frequency of 13.5%, an association p-value of 2.85x10⁻¹⁰² and 6.6Å from the RIF binding site, with a further 53 mutations close to the RNAP β-subunit active site. Both stabilising and destabilising mutations were observed. Most destabilising mutations can be attributed to changes in side chain charge, length and bulkiness. Of the most stabilising mutations, N409S and D407V lie within the active site. The latter possibly increases stability due to loss of negative charge. Located at the β’ subunit interface, H695Y mutation possibly forges internal interactions with residues, thereby supporting the reduced affinity for other subunits.
Figure 1. RNA polymerase β-subunit the target of RIF with effects of mutations highlighted for A with RIF docked and B DNA bound.

Similar observations were made for targets associated with INH and DCy. These valuable insights can be used to aid the development of future drugs and, via their integration into predictive tools, in pathogen surveillance.
Abstract

Residue Interaction Networks (RINs) have been used to study a plethora of phenomena. In a RIN nodes represent amino acids and the connections between them depict non-covalent interactions occurring in the three dimensional structure of proteins. Here we describe RIP-MD, a Visual Molecular Dynamics (VMD) plugin to facilitate the study of RINs using both static structures and trajectories obtained from Molecular Dynamics (MD) simulations of proteins.

Introduction

To perform their function, proteins adopt a three dimensional structure resulting from the sequence-based orchestration of both covalent and non-covalent interactions. The structure of a protein is not static; residues may undergo conformational rearrangements and, in doing so, creating, stabilizing or breaking non-covalent interactions. MD is a computational-based technique widely used to simulate these movements with atomic resolution. However, given the data-intensive nature of the technique, gathering relevant structural and functional information from MD simulations is a complex and time consuming process requiring several computational tools to perform these analyses. Among different approaches, the study of RINs has proven to facilitate the study of protein structures such as allosterism, enzymatic activity, protein folding and in protein-protein interactions.

Methods

The general workflow of RIP-MD is shown in Figure 1. RIP-MD starts with a dynamic (MD trajectory) or static (PDB) protein structure plus the parameters that define the interactions as input (Fig. 1.A). The next step pre-processes the input to ensure it
complies with the required format (Fig. 1.B). Following, a search for interactions between all atoms is carried out in each snapshot or the static structure (Fig. 1.C). In the last step (Fig. 1.D), RIP-MD generates the output files that include correlation maps and files that define a RIN that can be characterized in network visualization tools such as Cytoscape. Below, each of these steps is described in detail.

Figure 1: RIP-MD Workflow

Results & Conclusions

Our software generates RINs for static protein structures or from MD trajectory files. The non-covalent interactions defined by RIP-MD include Hydrogen bonds, Salt bridges, Van der Waals contacts, cation-π, π-π, Arginine-Arginine and Coulomb interactions. In addition, RIP-MD also computes interactions based on distances between Cαs and disulphide bridges.

Results are displayed using the VMD capacities, whereby through some steps, it is possible to select and visualize interactions described for residues in a MD trajectory. Network and descriptive table files are also generated, allowing their further study in other specialized platforms. Furthermore RIP-MD generates correlation plots, where relationships between the dynamic behavior of different parts of the protein can be determined and quantified.

As example cases we studied two different MD systems: a GJC complex and MD2. Regarding the GJC, a comparison of the initial structure and a short MD simulation revealed that inter-chain interfaces are stabilized mainly by HBs and SBs, and that Arg-Arg and Cation-π interactions tend to disappear over the trajectory. In the second case study, MD2, we focused on the closure process of this protein as reflected by Pearson correlation of HBs and VdWs interactions. Our study showed notable differences
between the different stages of the closure process and identified relationships between spatially separated residues which no direct interaction between them.

RIP-MD is available at http://www.dlab.cl/ripmd.

References


4. del Sol A., O'Meara P. Proteins, 58(3):672–82. 2005
31 - Combining Sequence and Structural features lead to accurate interpretation of genetic variation: large scale modeling and classification of de novo human mutations with VIPUR

Richard Bonneau¹,², Evan Baugh¹, Julia Koehler-Leman², Nicholas Carriero², Christian², Julia Koehler-Leman²

¹New York University, ²Simons Foundation

Interpreting coding variation (detecting mutations and small indels that disrupt function on a large background of neutral mutations) is a key challenge that spans several biological applications (from interpreting human variants to annotation of novel bacterial genomes). Widely used methods for interpreting protein variation typically focus on annotating mutation pathogenicity rather than detailed interpretation of variant deleteriousness and frequently use only sequence-based or structure-based information. We present our approach to this problem, VIPUR, a computational framework that seamlessly integrates sequence analysis and structural modeling (using the Rosetta protein modeling suite) to identify and interpret deleterious protein variants. We will describe our work to curate a very large database of biophysically characterized variants in several species. Using this large benchmark we have shown that VIPUR outperforms other methods for variant interpretation, that several widely used sequence-based methods perform significantly worse than advertised on de novo mutation, and that structural features account for much of this improvement. VIPUR can be applied to mutations in any organism’s proteome with improved generalized accuracy (AUROC .83) and interpretability (AUPR .87). We demonstrate that VIPUR’s predictions of deleteriousness match the biological phenotypes in ClinVar and provide a clear ranking of prediction confidence. We will discuss new transfer-learning based approaches to expanding VIPUR to allow for classification of variants in soluble, membrane and disordered regions in proteins, greatly expanding the reach of structural-based methods for interpreting coding variation. We will provide detailed examples of using VIPUR to interpret known mutations associated with inflammation, autism spectrum disorder, and diabetes.


PMID: 26926108; PMCID: PMC4824117.
Development of new Rosetta energy functions and rotamer libraries for modeling and design of hybrid systems containing proteins, non-canonical amino acids and peptidomimetic backbones

P. Douglas Renfrew², Richard Bonneau¹,²

¹New York University, ²Flatiron Institute, Simons Foundation

We will discuss new methods to incorporate non-canonical side chains, peptidomimetic backbones and other synthetic moieties into the Rosetta energy function. The PDB affords vast amounts of data for parameterizing and testing protein and nucleic acid scoring functions and several challenges arise when working (in Rosetta or other macromolecular modeling frameworks) with biomimetic oligomers and non-canonical side-chains. This work proceeds in three coordinated phases: 1) representation, including the creation of peptidomimetic and NCAA rotamer libraries, 2) scoring and energy function development, including the development of split reference energies and new work on additional energy components, and 3) testing the method, including applications of new hybrid protein-peptide-peptidomimetic design protocols to inhibiting biomedically relevant protein interfaces. We will first focus on resolving a key challenge for all design applications: deriving appropriate hybrid reference (unfolded/unbound) energies (a part of most macromolecular design frameworks that we like to ignore as much as possible). Algorithms for the design of biomacromolecules typically employ "reference energies" to correctly weight the probability of designing particular residue types and account for the intrinsic (unfolded) energy of key variable moieties like side-chains. For the canonical amino acids these quantities can be treated as free parameters and optimized such that designed sequences reproduce native sequence amino-acid frequencies (approaches not applicable for non-canonical amino acids (NCAAs) and rare/novel moieties). More general methods exist to produce reference energies for NCAAs that provide a proxy for the energy of each residue type in an "unfolded" state if the unfolded state can be explicitly modeled. We will describe a new two-component reference energy (TCRE) that combines a single-residue score terms (computed for each side chain) with an inter-residue score term that is an aggregate over atom-types present in each side-chain derived from high-quality protein structures. Our method does not require explicit modeling of the backbone/context of each moiety and avoids several problems we encounter when building these energy terms for non-canonical backbone chemistries like peptoids and beta-peptides. With this new reference energy integrated into our hybrid scoring function we improve performance over multiple hybrid design tasks and obtain improved sequence recovery. We will also
describe our work to build rotamer libraries for peptoids, OOPs, beta-peptides and L- and D-mixed peptides. After describing our work on the hybrid (protein plus peptidomimetic) scoring function and NCAA rotamer libraries we will describe recent results where we use these novel scoring methods to design peptidomimetic inhibitors of multiple biomedically relevant interfaces.


BioRiv preprint on beta-peptide rotamer libraries (accepted at Structure): http://biorxiv.org/content/early/2016/11/08/086389


33 - Unifying genomics and molecular biology data


GN7 of the Spanish National Institute for Bioinformatics (INB-ELIXIR) and Biocomputing Unit, National Center of Biotechnology (CSIC), C/ Darwin nº 3, Campus of Cantoblanco, 28049 Madrid, Spain.

With the advent of next generation sequencing methods, the amount of proteomic and genomic information is growing faster than ever. Several projects have been undertaken to annotate the genomes of most important organisms, including human. For example, the GENCODE project (Harrow, et al., 2012) seeks to enhance all human genes including protein-coding loci with alternatively spliced variants, non-coding loci and pseudogenes. Another example is the 1000 genomes (Genomes Project, et al., 2010), a repository of human genetic variations, including SNPs and structural variants, and their haplotype contexts. These projects feed most relevant biological databases as UNIPROT (2015) and ENSEMBL (Cunningham, et al., 2015), extending the amount of available annotation for genes and proteins.

In this work we present the a web platform, 3DBIONOTES-WS (Tabas-Madrid et al., 2016), that aims to merge the different levels of molecular biology information, including genomics, proteomics and interactomics data into a unique graphical environment. The first and current version of 3DBIONOTES integrates proteomic, genomic and functional annotations with structural data, providing a unified and interactive view of the different sources of information. Current development offers a unified view of three of the most relevant protein databases: UniProt (2015), PDB (Berman, et al., 2014) and EMDB (Lawson, et al., 2011), onto which other sources of biological annotations are also provided, such as PhosphoSitePlus (Hornbeck, et al., 2015), Immune Epitope DB (Vita, et al., 2015), BioMuta (Wu, et al., 2014) and dSysMap (Mosca, et al., 2015).

3DBIONOTES platform is comprised of the web server and client. The web server performs three major tasks. First, it maps the residue identifiers of PDB structures with the amino acid indexes of UniProt sequences The SIFTS (Structure Integration with Function, Taxonomy and Sequence) (Velankar, et al., 2013) resource provided by the EBI (European Bioinformatics Institute) is used to compute the mappings. Second, it provides the relation between EMDB maps, PDB structures and UniProt sequences. These relations are obtained accessing the different web services provided by the EBI (Meldal, et al., 2015). And third, finally, the server also supplies the protein
The web client provides an interactive environment connecting protein sequences, structures and annotations. The client comprises three major panels: the structural panel, the sequence panel and the annotation panel. All panels are interconnected, leading to graphic interactivity between them; thus, when a protein annotation from the annotation panel is clicked, the protein sequence region related to the annotation and the residues in the corresponding chain of the structural panel are highlighted. Additionally, when a segment of sequence is selected in the sequence panel, the annotation panel displays the selected region so that the particular annotations falling inside are marked and the corresponding residues of the structure are highlighted. 3DBIONOTES web application is accessible at http://3dbionotes.cnb.csic.es.


34 - On the turning away

Alexandre G. de Brevern

INSERM, Univ Paris Diderot, Sorbonne Paris Cite

Abstract

β-turn IV, *i.e.*, the miscellaneous category, represents near 1/3rd of β-turn residues in protein structure, and is the second most frequent β -turn. An innovative clustering approach was able to underline the existence of different new turns not previously described. The four most occurring clusters defined the new β-turn types. They exhibit interesting sequence – structure relationships.

Introduction

The classical secondary structures are composed of α-helices and β-strands connected by coil. Two other repetitive structures also exist, namely the PolyProline II and the β-turns. These last have been characterized by a hydrogen bond between N-H and C=O of residues *i* and *i+3* by Venkatachalam. He also characterized the first β-turn types. Later novel turns were defined, some being discarded, leading to a final collection of type I, I’, II, II’, IV, Vla1, Vla2, Vlb, and VIII β-turns. Turns that do not fit any of the above criteria are classified as type IV. β-turn IV, *i.e.*, the miscellaneous category, represents near 1/3rd of β-turn residues in protein structure, and is the second most frequent β -turn. An automatic clustering approach based on the rules of β-turn type assignment was designed to search for recurrent new turns inside this miscellaneous type. A comparison with related studies and amino acid sequence over- and underrepresentation was performed underlying interesting features.

Methods

Different non-redundant dataset of protein structures were taken from low to higher redundancy to analyse potential bias of the dataset (none was found). From these datasets, were taken only the type IV β-turns encoded as their series of central dihedral angles.
A specific clustering approach was designed to cluster type IV β-turns by using the classical rule, allowing +/- 30° for all angles, with the exception of one at +/-45° for the defined values. The clustering derived from Self-Organizing Maps (SOM, without diffusion between the clusters)\textsuperscript{4}. The training was carried out in 2 successive parts; the first one limited the potential bias of initialization, and the second refined the clustering by using the specific rules for β-turn types. The type IV β-turns were selected from a dataset D. Thus, each dataset was associated with one new type IV β-turn.

**Results & Conclusions**

![Figure 1: Ramachandran plot of β-turns. (Left) a close-up of type II and IV1 β-turns, and (Right) on type VIII and IV2 β-turns, the first square corresponds to the +/- 30° rule, and the second one to the +/- 45° rule.](image)

Surprisingly, with 10 different datasets, the unsupervised training was highly robust, producing always the same four major clusters (with negligible variation). These types, named IV\textsubscript{1}, IV\textsubscript{2}, IV\textsubscript{3} and IV\textsubscript{4}, represent half of the type IV β-turns, and are more frequent than many established ones. Figure 1 shows a direct comparison of the two most frequent new turns with their closest relatives underlying their relationship but also their differences. Type IV\textsubscript{1}, is in the neighbourhood of type II but with very different amino acid composition, while IV\textsubscript{2} is close to type VIII with related amino acid content. Types IV\textsubscript{3} and IV\textsubscript{4} are in the same dihedral angle region than frequent β-turn type I, but with distinct dihedral angle values.

Comparisons with the previous alternative classification proposed by Efimov \textsuperscript{5} and
Thornton\(^6\) emphasized the uniqueness of the approach. Notably, the most frequent new turn (type IV\(_1\) β-turn) was not highlighted, although it is the 5\(^{th}\) most occurring turn (including rest of type IV β-turns). Only the type IV\(_2\) β-turns were previously included.

References

35 - A complete Web resource for Galactosemia-related proteins

Bernardina Scafuri¹, Antonio d'Acierno¹, Angelo Facchiano¹, Anna Marabotti²

¹CNR-ISA, National Research Council, Institute of Food Science, Avellino, Italy, ²Dept. Chemistry and Biology “A. Zambelli”, University of Salerno, Italy

Abstract

We have developed a free, Web-accessible database in which we have collected information about the predicted structural and functional effects of missense mutations of the three enzymes of the Leloir pathway, whose impairment is linked to different forms of the genetic rare disease galactosemia.

Introduction

When the monosaccharide galactose is introduced in the human body, it is metabolized by a biochemical pathway, called Leloir pathway, involving three enzymes: Galactokinase (GALK, E.C. 2.7.1.6), which phosphorylates galactose to galactose-1-phosphate; Galactose-1-phosphate uridylyltransferase (GALT, E.C. 2.7.7.12), which converts galactose-1-phosphate into glucose-1-phosphate with the simultaneous epimerization of uridine diphosphate (UDP)-glucose into UDP-galactose; UDP-galactose-4’-epimerase (GALE, E.C. 5.1.3.2), which catalyzes the conversion of UDP-galactose into UDP-glucose.

Mutations in the genes coding for these three enzymes are associated to three different autosomal recessive deficiencies: classic galactosemia or type I galactosemia (OMIM #230400), associated to GALT deficiency; galactokinase deficiency or type II galactosemia (OMIM #230200), associated to GALK deficiency; galactose epimerase deficiency or type III galactosemia (OMIM #230350), associated to GALE deficiency. Each disease has different clinical manifestations with different severity, depending also on the mutation. The most common ones are missense mutations that can have different effects on the protein structure and function. We have performed a thorough study in order to predict these effects starting from the crystallographic structures of these enzymes.
Methods

The crystallographic structures of the three human enzymes of the Leloir pathway are available through the PDB database. Starting from these structures, we modelled each missense mutation associated in literature to galactosemia, and we evaluated their impact on different structural and functional features: secondary structures, solvent accessibility, intra- and intersubunit interactions, interactions with the substrate, protein stability, using online servers or well-known analysis tools installed locally. All the data obtained have been stored into a database developed using PostgreSQL, whereas the Web interface have been designed using JAVA as coding language and Apache Struts 2 as MVC framework. Netbeans has been used as development tool.

Results & Conclusions

The present resource supersedes two other resources previously developed in our laboratory: GALT Protein Database (http://bioinformatica.isa.cnr.it/GALT/GALT2.0/), focused only on GALT enzyme, and the preliminary version of Galactosemia Proteins Database 1.0. The differences with the present upgraded database rely both on the raw data and the analyses made. In the current database, more than 250 galactosemia-related missense mutants of GALT (either heterozygous or homozygous), 22 mutants of GALE and 15 mutants of GALK are stored. Most of the data collected have been obtained by using new tools that have been made available during the time, and/or developing ad hoc tools. The database is logically composed by four interconnected sections in order to manage information about the wild type and the mutant proteins, both in homozyogous and in heterozygous form. Finally, a Web interface has been developed with many filters, keeping in mind two opposite needs: on one hand, the possibility to interact with non-experienced users (including patients and their relatives) providing them useful information; on the other hand, the opportunity for expert researchers to gain full information about the predicted effects of mutations on the structural and functional features of these enzymes. This knowledge can indeed be helpful to better understand and, possibly, to predict, the clinical outcome of a mutation on individuals carrying it.

Acknowledgements

The project is funded by FARB-ORSA-143855. B.S. is supported by a Elixir fellowship.
36 - DisProt 7.0: a major update of the database of disordered proteins

Damiano Piovesan¹, Francesco Tabaro¹, Marco Necci¹, Silvio Tosatto¹,²

¹Department of Biomedical Sciences, University of Padua, Italy, ²CNR Institute of Neuroscience, Padua, Italy

Abstract

The Database of Protein Disorder (DisProt, URL: www.disprot.org) has been significantly updated and upgraded since its last major renewal in 2007. Information has been curated on more than 800 entries of intrinsically disordered (ID) proteins or regions that exist and function without a well-defined three-dimensional structure. We have re-curated previous entries to purge DisProt from conflicting cases, and upgraded the functional classification scheme to reflect continuous advance in the field in the past 10 years.

Introduction

The structural and functional characterization of disordered proteins represents a special challenge, because they exist as an ensemble of rapidly interconverting conformations. Although they cannot be crystallized, various techniques can report on their highly dynamic structural state at low- or even high spatial and temporal resolution. The Database of Protein Disorder (DisProt, URL: www.disprot.org)¹ is the major repository for manually curated ID annotation. Here, we present the first major update made to DisProt (version 7.0) released in the last couple of years². DisProt 7.0 presents classified knowledge regarding the experimental characterization and functional annotations of ID, and is intended to provide an invaluable resource for the scientific community for a better understanding intrinsic disorder and the development of better computational tools for ID proteins.

Methods

DisProt database records are of two types, ‘protein’ including general information about the protein and ‘disordered region (DR)’ including evidence of disorder from literature. DR records are evidence-centric, i.e. different documents are stored for different experiments even when related to the same region allowing to track annotation evidence
unambiguously. DR records also include experimental evidence quality tags for ambiguous annotations (engineered sequences, fragments, unclear disorder boundaries, non-physiological conditions). The major improvement from previous versions is the manually curated functional annotation of the regions. Whenever possible, curators-associated functions based on literature evidence are indicated by selecting terms from a new ontology split into three parts. (i) Molecular function of disorder describes the type of functional readout of function (such as molecular chaperone). (ii) Molecular transition necessary for function (such as disorder-to-order transition). (iii) Molecular partner that is recognized by the disordered protein (such as protein/RNA/DNA/small molecule).

Results & Conclusions

DisProt content was expanded by defining a standardized set of experimental techniques and a novel functional ontology of disordered segments. The other main improvement in DisProt is a complete re-annotation of existing entries to remove inconsistencies and an expansion of ca. 50% over the previous release, which also resulted in a significant shift in the length coverage of disordered regions in the database (Figure 1). This advance was made possible by a distributed annotation effort coordinated by the COST Action NGP-net (URL: ngp-net.bio.unipd.it). Finally, we hope that the upgrade of DisProt will encourage the scientific community to deposit experimental evidence for disorder, and that this renewed momentum will lead to an increased awareness of the importance of intrinsic disorder in proteins.
FIGURE 1. Distribution of disorder segment lengths. The current DisProt release is distinguished by experimental technique, the previous (gray) did not have this information in a machine-readable format.

References


37 - 40-fold increase in coverage of structure-based annotations for UniProt entries via the SIFTS resource

Jose M. Dana¹, Guoying Qi², Nidhi Tyagi², Aleksandras Gutmanas¹, Claire O'Donovan², Maria Martin², Sameer Velankar¹

¹Protein Data Bank in Europe (PDBe) / EMBL-EBI, ²UniProt / EMBL-EBI

The Structure Integration with Function, Taxonomy and Sequences resource (SIFTS; PDBe.org/SIFTS) was established in 2002 and continues to operate as a collaboration between the Protein Data Bank in Europe (PDBe; pdbe.org) and UniProt (uniprot.org). The resource is instrumental in the transfer of structure based annotations for protein sequence data through provision of up-to-date residue-level mappings between entries from the PDB and from UniProt. SIFTS also provides residue-level annotations from other biological resources: currently IntEnz, GO, Pfam, InterPro, SCOP, CATH, PubMed and the NCBI taxonomy database.

The information is made available to the scientific community in different formats such as individual XML files for each PDB entry, or CSV and TSV files for the whole PDB archive. The PDBe REST API (PDBe.org/api) allows for programmatic access to SIFTS. Many bioinformatics resources (e.g. UniProt, RCSB, Pfam, etc.) use SIFTS data to obtain cross-references between the PDB and other biological databases in order to serve their users with up-to-date information.

Until 2017, SIFTS mappings between PDB and UniProt were always calculated with respect to the canonical sequences of proteins. A number of improvements were recently carried out by PDBe enabling overlapping mappings and, as a result, SIFTS now supports mappings to isoform sequences and to UniRef90 clusters (all sequences in UniProt which have 90% or greater sequence identity).

More specifically, all the isoforms of canonical UniProt accessions are now mapped, allowing a better representation of PDB structures that are representative of a specific isoform (e.g. PDBe.org/1loi).
Extending the mapping to UniRef90 cluster members expands the structural coverage of UniProt almost 40-fold. From ~40000 UniProt accessions mapped directly to PDB entries to more than 1.5 million UniProt accessions with at least 90% sequence identity and 70% sequence coverage by the structures in the PDB. Specifically, our analysis shows that while the PDB contains fewer than 3000 unique human proteins, there are a further 3300 proteins from other organisms, for which there is at least one structure in the PDB and which are highly similar, at the sequence level, to human proteins otherwise not available in the PDB.

Future work will include Pfam annotation updates “in-between” releases (providing more up-to-date information) and new residue-level annotations for Ensembl, COG and OrthoDB.
Alzheimer is a neurodegenerative pathology where the formation of Aβ plaques gives rise to neurons death. Recent studies have suggested a role of GPR3 receptor in the formation of Aβ plaques. So far only one synthetic agonist, Diphenyleneiodonium (DPI), is known for this receptor. Here, using state-of-the-art bioinformatics approaches, we investigated the binding cavity of GPR3 and different putative interactions with DPI. Our predictions were validated by site-directed mutagenesis.

Introduction

Alzheimer is a neurodegenerative disease characterized by loss of brain connectivity, leading to the deprivation of memory[1]. Lately, it was discovered that the orphan and constitutively active G-protein coupled receptor 3 (GPR3), belonging to Class-A GPCRs, is involved in Alzheimer’s disease (AD), through direct interaction with β-arrestin 2[2]. Although GPR3 is considered a new promising therapeutic target for AD, it has not a solved structure and a known binding site, yet. Here, using bioinformatics and computational biophysics techniques we characterized the interaction of GPR3 with its synthetic ligand, DPI. Our predictions were validated by using site-directed mutagenesis experiments. We have also identified a mutant that completely abolishes the basal activity of the receptor by modifying the allosteric sodium (Na+) putative binding cavity. The latter could help the determinants’ characterization underlying the β-arrestin pathway, directly correlated with the plaques formation.

Methods

We modeled GPR3 and predicted its binding cavity by using the GOMoDo[3] web-server. The obtained model was manually checked to inspect the conservation of the class A GPCRs structural fingerprints. DPI were docked into the modeled GPR3 by
using the Haddock[4] server on GOMoDo. The best complex structure, i.e. the one showing the lowest energy of the most populated cluster was considered as input for Molecular Dynamics simulations (MDs). The 200 ns MDs were performed by using the molecular mechanics/coarse-grained (MM/CG) hybrid approach[5], developed in our laboratory. In the hybrid simulations, the ligand, the putative binding site and a drop of water molecules are included in the MM region and the rest of the receptor is treated as CG (Figure 1). Site-directed mutagenesis were carried out with the ‘QuickChange Site Directed Mutagenesis’ kit.

Results & Conclusions

The model was obtained based on the structure of human β2 adrenergic receptor as template, sharing ~23% of sequence identity. The best docking GPR3-DPI complex was funneled to hybrid MM/CG simulations approach to exhaustively explore the conformational space of the ligand, the binding cavity and the hydration shell. Our simulation results pointed to a network of residues putatively involved in ligand interactions (Figure 1).

Figure 1. GPR3 in complex with its agonist DPI.
Our model was validated by performing wet-lab alanine scanning mutagenesis on residues Cys267 and Phe120, putatively involved in halogen and π-stacking interaction with the ligand, respectively. Conversely to the previous two mutants, D55A, putatively involved in the allosteric Na+ binding cavity, completely abolished the cAMP signal[6]. Previous studies have shown that mutants localized in this binding cavity biased receptor towards active conformations, increasing the cAMP or Ca+ signal. However, GPR3 is a constitutively active receptor and D55A could promote inactive conformations of the receptor and biased towards the β-arrestin pathway. Experiments are carrying out to validate our hypothesis.

References


6. Katritch et al. TBS
39 - Unraveling the histone code by fragment blind docking

Csaba Hetenyi¹, Nikolett Meszaros², Istvan Horvath³, Monika Balint¹

¹Department of Pharmacology and Pharmacotherapy, University of Pecs, Hungary.,
²Department of Biochemistry, Eotvos University, Hungary., ³Chemistry Doctoral School, University of Szeged, Hungary.

Abstract

Histones are key actors in pathomechanisms of epigenetics-based diseases. Due to combinations of post-translational modifications of their N-terminal peptide tails, they can form various complexes with partner proteins. Determination of the structures of the complexes would be a key to unravel the histone code and the design of new drugs. The large number of possible complexes imposes a true challenge for experimental structure determination techniques. To accelerate the work in this field, a fragment blind docking approach is presented for prediction of structure of histone peptides bound to their protein targets.

Introduction

Epigenetic changes play a key role in pathomechanisms of various diseases and are often based on post-translational modification (PTM) of the N-terminal peptide tails of histone sub-types H3 and H4. PTMs include methylation, acetylation, phosphorylation, etc. of side-chains of certain amino acids (K, S, R, T, Fig. 1a)¹. Reader proteins are involved in “de-coding and translating” of information stored as a combination of the above PTMs. Thus, it was proposed¹ that beyond the genetic code, these combinations can be considered as an epigenetic “histone code”. Understanding the histone code by prediction of the structure of histone-reader (writer, eraser) complexes² is crucial for intervening in the pathological processes of epigenetic diseases. However, the determination of numerous atomic resolution structures of the above complexes is not trivial even with high throughput X-ray crystallography, if considering the large number of possible combinations of PTMs.

Methods
The blind docking$^3$ (BD) approach was applied to generate the structures of ligand molecules bound to protein targets, and no prior information on the location of the binding site was used. BD was performed with the popular program package AutoDock 4.2.3, and the search space was extended to the entire surface of the target. Before BD, all fragments were chemically blocked at the covalent bonds broken during fragmentation. After BD, molecular dynamics simulations were also performed as described in our previous study on a histone complex$^4$.

Results & Conclusions

The fragment docking (FD) strategy have been extensively used for docking studies focusing the search to a binding site (restricted search space). However, its use for BD has not been introduced. Here, we report on a combination of the two strategies (FD+BD) into a hybrid approach of fragment blind docking (FBD). Our preliminary results show that it was possible to generate a good collection of docked fragments during exhaustive BD of the N-terminal peptide fragments of the H3 peptide tail (Fig. 1bc). The FBD method is a promising alternative for rapid generation of protein-bound histone peptide structures, and it will help unraveling the “histone code”, an important step in epigenetic drug discovery.
Figure 1  a) N-terminal sequence of the histone H3 peptide tail. Amino acids prone to PTM are shown. b) An example of docked conformations of blocked dipeptide fragments of the tetrapeptide sequence highlighted with blue in a). Observe the empty space instead of the connecting amide bond between R and T. c) As in b) shown together with the crystallographic reference conformation (grey C atoms) of the full ARTK. A good structural match can be observed between the full tetrapeptide and its dipeptide fragments. Notably, the similarity between the blocking acetyl group and side chain of T resulted in the exchange of the positions of two groups by rotation during docking. The FBD algorithm corrects such cases easily by a single rotation (arrow). The protein target is not shown for simplicity’s sake.

References


40 - Interacting residues prediction based on a random forest multi-step approach

R. Sanchez-Garcia¹, J. Segura¹, C.O.S. Sorzano¹, J.M. Carazo¹

¹N7 of the Spanish National Institute for Bioinformatics (INB) and Biocomputing Unit, National Center of Biotechnology (CSIC), C/ Darwin no 3, Campus of Cantoblanco, 28049 Madrid, Spain.

Abstract

In this work we present a new machine learning-based method developed to predict residue-residue contacts of two proteins that are known to interact. Our approach is based on a three-step random forest classifier trained on both sequential and structural features. Performance evaluation was carried out by cross validation over the targets of Protein-Protein Docking Benchmark 5.0. When compared to related approaches, our method showed better than state-of-the-art performance with better running time.

Introduction

Protein-Protein Interactions (PPI) are essential for most cellular processes. In spite of their importance, just a few methods have been design to predict contacts at the residue-residue level. Those predictions, when accurate enough, can be employed as distance constrains for quaternary structure modeling, simplifying both model construction and validation. State-of-the-art techniques aimed to predict residue-residue contacts include correlated mutations-based methods [3] and machine learning-based approaches [1,2]. Here we present a new machine learning method developed to obtain residue-residue contact prediction from structural and sequential features.

Methods
Data: Our dataset was collected from the PDB files that constitute Docking Benchmark 5.0 [4]. Pairs of residues were labeled as interacting if the distance between any of their heavy atoms were less than 6.0 Å and as non-interacting otherwise. Each pair of amino acids was encoded by combining structural and sequential properties. Among structural features, accessible surface area, protrusion index and atomic depth stand out, whereas the main sequential features we have employed are Position Specific Scoring Matrices computed with psi-BLAST [5].

Classifier: We have trained serially three random forest classifiers. In the first step, we compute a residue-residue contact score using a classifier that has been trained over a data set where each residue pair has been codified employing features of the residues and of their sequential neighbors. In the second step, each residue pair is codified using the first step scores combined with features of the residues and their structural neighbors. Finally, the last step combines both first and second step predictions together with a selection of features of the residues and their structural neighbors.

Evaluation: We have evaluate the performance of our method by leave one complex out cross-validation over the PDB files used for training. ROC auc and precision/recall at different thresholds were computed for predictions at both residue-residue level and binding site level.

Results & Conclusions

On average, our predictions obtained ROC auc values slightly above 90% for pair prediction and above 80% for binding site prediction. When compared to similar approaches (table 1), our method offers the best performance for the same data set that was employed in those methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROC auc at pair level</th>
<th>ROC auc at binding site level</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPIPPE [1]</td>
<td>72.9</td>
<td>66.1</td>
</tr>
<tr>
<td>PAIRpred [2]</td>
<td>88.7</td>
<td>77.0</td>
</tr>
<tr>
<td>Ours</td>
<td>90.2</td>
<td>81.0</td>
</tr>
</tbody>
</table>

Table 1. Performance of different methods measured as ROC auc (percent). Evaluation was carried out using the same PDB targets for all methods.
In addition to better predictions, our method is also the fastest for both training and testing. Indeed, that speed improvement makes it possible to scale up our approach by using bigger training sets. As a consequence, better results may be expected for a future version.

References

41 - Improved Rosetta protein structure prediction with customised fragments libraries based on structural class annotations

Jad Abbass1,2, Jean-Christophe Nebel1

1Kingston University, 2Lebanese International University

Introduction

Since experimental techniques are time and cost consuming, in silico protein structure prediction is essential to produce conformations of protein targets. When homologous structures are not available, fragment-based protein structure prediction has become the approach of choice. However, it still has many issues including poor performance when targets’ lengths are above 100 residues, excessive running times and sub-optimal energy functions. Taking advantage of the reliable performance of structural class prediction software, it is proposed to address some of the limitations of fragment-based methods by integrating structural constraints in their fragment selection process.

Methods

Production of accurate predictions fragment-based protein structure prediction approaches is eased if, for each given position, there is high proportion of fragments fitting closely the native one [1]: the higher the quality of the fragment libraries, the more focus the conformation search is on the sub-space containing the native structure. This property is exploited by customising further fragment libraries according to the nature of the protein target. More specifically, the set of template proteins which are the source of those libraries is tailored to the target so that their quality is increased. Since sequence based structural class prediction has become relatively mature, such information is used to select the relevant template structures.

From those principles, a new fragment-based protein structure prediction methodology has been designed [2]. First, structural class is predicted from the sequence of the protein target. Second, a target specific list of template structures is generated by extracting high resolution templates sharing the same structural class from the default template protein set (a PDB subset) associated to the fragment-based method. Finally, the target sequence and its associated template list are submitted to a fragment-based
protein structure prediction, which produces customised fragment libraries and generates a set of putative structures of the protein target.

Results & Conclusions

Using Rosetta [3], a state-of-the-art fragment-based protein structure prediction package, the proposed pipeline is evaluated on 70 former CASP targets containing up to 150 amino acids. Using CATH-based structural class annotations, enhancement of structure prediction performance is highly significant in terms of both GDT_TS (at least +2.6, p-values < 0.0005) and RMSD (~0.4, p-values < 0.005). Further analysis also shows that methods relying on class-based fragments produce conformations which are more relevant to user and converge quicker towards the best model as estimated by GDT_TS (up to 10% in average). This substantiates the hypothesis that usage of structurally relevant templates conducts to not only reducing the size of the conformation space to be explored, but also focusing on a more relevant area.

Experiments conducted during CASP11 confirmed those results: out of 12 targets, "Rosetta_at_Kingston" – our group – was better than "Baker" in 6 and better than "BAKER-ROSETTASERVER" in 6 in terms of GDT_TS as shown in Figure 1. (Note that some targets were divided into domains by CASP)

![Figure 1: Comparison of results of 12 targets amongst our group and the two formal groups of Rosetta team.](image)

References
44 - PERFORMANCE OF MACHINE-LEARNING SCORING FUNCTIONS IN STRUCTURE-BASED VIRTUAL SCREENING

Maciej Wójcikowski\textsuperscript{1}, Pedro Ballester\textsuperscript{2}, Pawel Siedlecki\textsuperscript{1}

\textsuperscript{1}Institute of Biochemistry and Biophysics PAS, Pawinskiego 5a, 02-106 Warsaw, Poland., \textsuperscript{2}Centre de Recherche en Cancérologie de Marseille (CRCM), Inserm, U1068, Marseille, F-13009, France.

\textit{Correspondence to pedro.ballester@inserm.fr or pawel@ibb.waw.pl}

**Introduction**

Scoring Functions (SFs) aim at approximating the relationship between the variables that characterise the X-ray crystal structure of a protein-ligand complex and its binding affinity. Whereas classical SFs assume a linear relationship, machine-learning SFs were introduced to employ more flexible models like Random Forest (RF)\textsuperscript{1}. As discussed in a recent review\textsuperscript{2} of the growing number of studies in this area, machine-learning SFs have shown great promise at predicting binding affinity and virtual screening (VS).

In the presented work\textsuperscript{5}, we investigate how including negative data instances (i.e. inactive molecules docked to targets) improves the performance of the machine learning SF. Such chimeric complexes are typically discarded from training procedures.

**Methods**

Here we constructed a new machine-learning SFs (RF-Score-VS v1, v2 and v3) trained on 15 426 active and 893 897 inactive molecules docked to a set of 102 targets. We use the full DUD-E data sets along with three docking tools, five classical and three machine-learning SFs for model building and performance assessment. Importantly, we follow validation practices that are more rigorous than those of most classical SFs\textsuperscript{2} (e.g. using cross-validation or not tuning model hyperparameters directly on the test set).
Results & Conclusions

Our results show RF-Score-VS can substantially improve VS performance: RF-Score-VS top 1% provides 55.6% hit rate, whereas that of Vina only 16.2% (for smaller percent the difference is even more encouraging: RF-Score-VS top 0.1% achieves 88.6% hit rate for 27.5% using Vina). In addition, RF-Score-VS provides much better prediction of measured binding affinity than Vina (Pearson correlation of 0.56 and −0.18, respectively). These experiments show that previously implemented machine-learning SFs for binding affinity prediction\textsuperscript{1,3,4} can excel at VS too, if appropriate care is taken. In addition, we further validate RF-Score-VS on an independent test set from the DEKOIS benchmark and observed comparable results.

![FIGURE 1. Predicted performance of classical vs machine-learning scoring functions for structure-based Virtual Screening. Top 1% of docked compounds predicted to be active across DUD-E targets by (A) Autodock Vina and its native SF (Rp = −0.18) and (B) RF-Score-VS v2 trained on horizontally split dataset (Rp = 0.56). Red points represent inactive decoys (false positives), whereas green points correspond to ligands (true positives). Predicted values by RF-Score-VS v2 come from independent cross-validation folds.](image)

Research on the optimal application of machine learning to structure-based VS is highly promising, but it is still in its infancy due to being a more complex endeavour than binding affinity prediction from crystal structure of protein-ligand complexes. Indeed, training data sets for structure-based VS are formed by active and inactive molecules docked to a range of targets. Therefore, these data sets are much larger than the crystal structures used binding affinity prediction and require prior docking of each considered molecule.
We are providing the code of the ready-to-use RF-Score-VS SFs (http://github.com/oddt/rfscorevs_binary). We also provide the code to build these machine-learning SFs (http://github.com/oddt/rfscorevs) to facilitate further research in this exciting area.

References


45 - Mining Functionally Conserved Building Blocks in Proteins

Florian Kaiser\textsuperscript{1,2}, Sebastian Salentin\textsuperscript{2}, Michael Schroeder\textsuperscript{2}, Dirk Labudde\textsuperscript{1}

\textsuperscript{1}University of Applied Sciences Mittweida, \textsuperscript{2}TU Dresden

Abstract

Proteins with shared catalysis mechanisms, ligand binding, or fold have often evolved to retain similar interaction patterns to conserve functionality. While some of these patterns are reflected by shared sequence motifs, distant proteins may use only a small set of similar structural building blocks.

By unifying geometrical and molecular interaction patterns in a mining algorithm, the presented approach has the potential to identify molecular building blocks with functional conservation and can be applied in protein function prediction and drug target screening.

Introduction

Long-range contacts of residues are often responsible to retain the fold and function of proteins. These building blocks can be of very limited size compared to the overall protein, but are nevertheless known to be indispensable\textsuperscript{1}. However, as the driving force of evolution is the conservation of function – rather than specific amino acid sequence – such building blocks are hard to grasp by multiple sequence alignment techniques.

Functional conservation depends on the ability to form valid ligand interactions and residue contacts to restrict motion of e.g. catalytic sites\textsuperscript{2}. Noncovalent inter-residue contacts are essential to preserve an ordered, and consequently functional, state of the protein. Using interaction features overcomes limitations of purely geometrical approaches, which might overlook functional conservations.

De novo high-throughput annotation of functionally conserved molecular building blocks could greatly contribute to the understanding of protein fold, function, or ancestry.

Results & Discussion
Based on the extension of earlier work, we present a novel workflow for automatic discovery of molecular building blocks without any a priori knowledge. We applied our method to plastocyanin and trypsin structures and were able to recover several conserved, sequence-separated, and family-specific building blocks (Figure 1).

With the unique combination of accurate substructure matching and comprehensive characterization of inter-residue contacts with the Protein-Ligand Interaction Profiler (PLIP), we envision the possibility to derive family-specific libraries of functionally conserved building blocks to classify protein family association.

Furthermore, this data might allow functional characterization of new protein structures and identification of similar ligand binding environments for drug target screening.

Methods

An extension of itemset mining for protein structure data was used to carve out similar geometries and long-range contacts of molecular building blocks. The interaction data was obtained by PLIP version 1.3.4 using default parameters and the command line flag for detection of intra-chain interactions. Predicted contacts were abstracted as pseudoatoms by calculating the midpoint between interacting atoms. The nonredundant
(BLAST $p$-value $10^{-80}$) datasets of plastocyanin (PF00127) and tyrosin (PF00089) contained 44 and 33 single chain structures and were derived from the Pfam\textsuperscript{7} database.

**References**

46 - Unveiling the inhibition mechanism of HIF-2α:ARNT dimerization by protein dynamics

Stefano Motta¹, Claudia Minici², Laura Bonati¹, Alessandro Pandini³

¹Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan, Italy, ²Department of Immunology, Transplantation, and Infectious Diseases, DIBIT Fondazione San Raffaele, Milan, Italy, ³Department of Computer Science, Brunel University London, Uxbridge, United Kingdom

Abstract

The hypoxia-inducible factor (HIF) mediates cellular responses to low oxygen stress through dimerization with the Aryl hydrocarbon Receptor Nuclear Translocator (ARNT) and DNA binding. Since HIF is critically important for the sustained growth and metastasis of solid tumours in humans, its targeted inhibition by means of small molecules could be used for cancer treatment. Here we shed light into the inhibition mechanism at a molecular level by comparison of dynamics and energetics of the HIF-2α:ARNT dimer in the unbound and inhibitor-bound form.

Introduction

HIF is a transcription factor belonging to the bHLH-PAS protein family with a role in sensing oxygen levels in the cell. Under hypoxia its oxygen-dependent degradation pathway is blocked and HIF translocates into the nucleus, dimerizes with ARNT and becomes transcriptionally active. Due to the common hypoxic environment of tumours, HIF inhibition has been proposed as a promising anticancer strategy and a druggable cavity was recently detected within the HIF-2α PAS-B domain¹. In 2015 X-ray structures were released for the HIF-2α:ARNT dimer in the unbound (PDB ID: 4ZP4), the 0X3-inhibitor-bound (4ZQD) and DNA-bound (4ZPK) forms². It was suggested that inhibition could involve perturbations of the dimer architecture through interdomain communication across protein-protein interaction interfaces, but unveiling the molecular details of this mechanism requires an investigation of the protein dynamics.

Methods

We used a combination of surface conservation analysis, Molecular Dynamics (MD) simulations, energetic and dynamic network analysis to study the differences in the
comparison of HIF-2α:ARNT unbound and 0X3-bound states. Functionally conserved surface residues were detected with ConSurf3. MD simulations (3x300 ns) were performed with GROMACS 5.1 using the Amber ff99sb-ildn* force field. Free energies of dimerization were estimated with the MM-GBSA method4 using a per-residue decomposition analysis. Residue correlated motions were detected using the dynamical cross-correlation network analysis5 and interdomain communication was recovered by optimal path analysis using Bio3D6.

Results & Conclusions

Functionally important regions consistently showed highly conserved residues: on the dimerization interfaces, in one of HIF-2α PAS-A loop, known to be involved in DNA binding, and in one of ARNT PAS-A loop, which role in enhancing dimerization was confirmed by our binding free energy decomposition analysis. MD analysis of the dimer in the unbound form revealed internal rigidity of the PAS domains with quaternary structure oscillations involving flexibility of the interdomain interfaces. Comparison with the 0X3-bound simulations highlighted a perturbed PAS-B:PAS-B interface, with a reduced binding free energy contribution in the bound form. Furthermore, a rigidification of one HIF-2α PAS-B strand was observed in the inhibited state. The analysis of correlated motions and interdomain communication (see Figure 1) unveiled the long-distance effects of this local perturbation, that decouples the dynamics of ARNT and HIF-2α PAS-A domains. Since the PAS-A:PAS-A interface is the primary dimerization region, this perturbation reduces the overall dimer stability. The here presented residue characterisation of this allosteric inhibitory mechanism will facilitate dynamics-based drug design.
FIGURE 1. Comparison of unbound and inhibitor-bound dynamic cross correlation network at the PAS-A:PAS-A interface. In the 3D view red edges represent the most perturbed correlated motions derived from the full correlation matrix (on the right).

References
47 - Deep Learning strategy for Improving ranking of protein fold recognition method ORION

Charlotte Perin$^{1,2,3}$, Yassine Ghouzam$^{1,2,3}$, Guillaume Postic$^{1,2,3}$, Jean-Christophe Gelly$^{1,2,3}$

$^1$University Paris Diderot - Paris 7, $^2$INSERM, $^3$INTS

Abstract

Ranking of potential interesting template is a key issue for comparative modelling based protein structure prediction. We have introduced a machine learning strategy based on Deep Learning to improve fold recognition performance ranking and detection of targets.

Introduction

Fold recognition is crucial for predicting protein structure by comparative modelling. We have recently developed ORION [1,3], a fold recognition method based on an original pairwise alignment of hybrid profiles between a query protein and a target template protein. Hybrid profiles combine both sequence and structure information for a more sensitive search which increases sensitivity for remote homology detection. ORION [1], encodes local structural information using a structural alphabet named Protein Blocks [3]. In the second version of ORION [2], solvent accessibility as an additional structural descriptor has been introduced. With the last version of ORION, we yielded 57% of true positive in Top 5 compared to the 11% and 29% for PSI-BLAST [4] and state-of-the-art HHsearch [5], respectively, in terms of detection sensitivity at fold level on the HOMSTRAD benchmark [6].

However, we have noticed that promising targets are not well ranked by ORION. Deep Learning, a novel and powerful Machine learning approach, is a promising approach to enhance performance of our method.

Methods

Recently we introduced in ORION a new novelty to improve its detection and ranking. The purpose is to re-rank all the targets from the ORION output in order to obtain a better correlation between the ranking and the true positive targets. Our strategy relies
on Deep Learning [7], a recent and powerful machine learning algorithm which has met striking successes in many applications such as detection of breast cancer in histology images [8], drug toxicity prediction [9] or game Go [10].

Results & Conclusion

Combining new descriptors and ORION output, Deep Learning approach is able to increase performance of more than 10% for the Top 5 targets compared to ORION alone.

References

48 - Molecular Dynamics Simulation of a 17β-Estradiol Specific DNA Aptamer

Alexander Eisold¹,², Dirk Labudde¹

¹University of Applied Sciences Mittweida, ²Institute for Organic Chemistry, Technische Universität Bergakademie Freiberg

Abstract

We present an approach that uses a generated three-dimensional (3D) model of an aptamer in order to describe its interactions with 17β-estradiol (E2) through MD simulations. This procedure allowed us to determine that between an aptamer’s asymmetric interior loop and E2 a hydrogen bond and a π-stacking interaction are formed. Both play a significant role in the stable bound state and allow a better understanding of the aptamer-E2 complex.

Introduction

Aptamers are short single-stranded oligonucleotides that fold into a complex 3D structure in vivo, which can bind molecules specifically¹. This specificity makes them suitable for detection of various targets at low concentrations in different environments. Hence, aptamers are suitable biosensors to detect substances such as E2²,³. E2 is the most potent endogenous steroid estrogenic sex hormone in mammals⁴ and because of this, exogenous E2 acts as an endocrine disruptor and has damaging effects on human male reproductive health⁵,⁶. This is a concern, since E2 is present at low concentrations in ground and drinking water⁷, which requires its detection in environmental samples⁵. Alsager and collaborators developed a 35-mer DNA aptamer with high binding affinity to E2 in 2015³. Its structure and binding processes have not been described yet. Resolving these two aspects is necessary for a generalized understanding of the interactions between an aptamer and similar target families. Our study focuses on two aspects: (i) development of a 3D-model of the above mentioned 35-mer DNA aptamer through secondary structure prediction (SSP) and coarse-grained modeling, and (ii) identification of the binding site through molecular dynamics (MD) simulations of the generated aptamer structure, with E2 as ligand.

Methods
The DNA hybridization determined by SSP\textsuperscript{8} was used as a basis for the generation of the aptamer 3D-model by coarse-grained modeling\textsuperscript{9}. The 3D-model with the lowest free energy was selected as starting point for MD simulations. The simulation (aptamer and E2 ligand) was carried out with GROMACS\textsuperscript{10} using explicit solvent, standard temperature and pressure conditions, and the AMBER99 force field\textsuperscript{11} for 25 ns.

Results & Conclusions

We found that E2 binds to an asymmetric interior loop of the 35-mer DNA aptamer formed by T12 and G23 to T25 (see figure 1A). In this region, E2 establish a hydrogen bond between its hydroxyl group at 3-β-position as donortype and the negatively charged oxygen from the phosphate of C26 backbone atom as acceptor. Additionally, it forms a parallel π-stacking interaction between E2’s cyclohexane ring at 3-β-position and the pyrimidine ring of T24 (see figure 1B).

We found that π-stacking interactions between E2 and the aptamer binding site play a role as significant as that of hydrogen bond during the binding process and the stable bound state.

FIGURE 1. (A) Modeling structure of the anti-E2 35-mer aptamer obtained from the 25 ns MD simulation (without water or ions). (B) Aptamer binding site and E2 ligand with highlighted different interactions.

References
49 - How is structural divergence related to evolutionary information?

Diego Javier Zea¹, Alexander Monzon², Gustavo Parisi², Cristina Marino-Buslje¹

¹Fundación Instituto Leloir, ²Universidad Nacional de Quilmes

Abstract

Structural divergence is high in protein families. What's more, residue contacts and solvent accessibility change a lot between the structures of a protein family. Consequently, the correlation value between conservation and solvent accessibility depends on the selected structure. Covariation scores give high scores to positions that are in contact in all the structures.

Introduction

A mutation in a biologically important position of a protein is either eliminated by purifying selection or compensated by mutations in other positions. These evolutionary events leave a signal that can be measured by conservation and covariation, respectively, in a Multiple Sequence Alignment (MSA).

Previous studies linking evolutionary and structural information, used a single structure as a representation of the structural space of a given protein family (represented in an MSA). This approach was applied in the optimization of covariation method parameters to predict residue contacts [1]. It was also applied in studies focused on understanding the structural constraints of protein evolution [2]. In particular, Relative solvent Accessible Surface Area (RASA) and Contact Number (CN) were identified as the main determinants of the evolutionary rate [3,4].

In this work we studied how evolutionary information is related to structural information when considering all the available homologous structures. In particular, we studied the relationship between conservation and RASA and also the performance of covariation methods for contact prediction when different structures of an MSA are used. It is worth noting that coevolutionary analysis requires MSAs populated by a huge number of homologous sequences, so a large structural divergence is expected.
Methods

We performed the analysis using ad hoc scripts in Julia language, taking advantage of the MIToS toolkit [1]. We selected Pfam families (v30.0) with at least 4 sequence clusters at 62% identity with structural information. Structures should cover at least 80% of the MSA columns and be longer than 30 residues. Finally, the dataset was randomly split into an exploratory (245 families) and a confirmatory (572 families) set in order to avoid the post hoc theorizing.

We measured conservation as Kullback-Leibler divergence and Shannon entropy. Corrected Mutual Information (ZMIp)[1] and Gaussian DCA[5] between column pairs were measured as a proxy to coevolution.

Results & Conclusions

As expected, Pfam alignments show a large structural divergence. We found that on average ~51% of the MSA columns change its exposed/buried status on average. Furthermore, on average ~53% of the residue contacts between column pairs that are present in at least one structure, are not in contact in another structure (changing contacts).

The correlation coefficients between RASA and conservation change when different structures are used. The average of the RASA values from multiple PDB residues found in an MSA column explains more conservation variance than RASA values from a single structure. Moreover, in our dataset RASA correlates better with conservation than CN or flexibility.

We saw that the predictive performance of covariation scores does not depend on the selected structure. This is because covariation measures tend to give a higher score to positions pairs that are in contact in all the known structures (conserved contacts).
Figure 1. Area Under the ROC Curve (AUC) of ZMlp and Gaussian DCA to predict contacts that change between structures of the same MSA versus AUC to predict conserved contacts.

References


50 - New binding site of the quorum sensing molecule N-3-Oxododecanoyl Homoserine Lactone with the transcriptional regulator LasR of Pseudomonas aeruginosa: Insights from Molecular Docking and Dynamics Simulations

Hovakim Grabski¹, Lernik Hunanyan¹, Susanna Tiratsuyan¹,², Hrachik Vardapetyan¹

¹Russian-Armenian University, ²Yerevan State University

Abstract

In 2017 World Health Organization announced the list of the most dangerous superbugs and among them is <em>Pseudomonas aeruginosa</em>, an antibiotic resistant opportunistic human pathogen. This organism attacks human patients suffering from such diseases as AIDS, cancer, cystic fibrosis, etc. Current therapies lack efficacy, partly because <em>P. aeruginosa</em> creates and inhabits surface-associated biofilms conferring increased resistance to antibiotics and host immune responses. However, there remains a need for understanding of the LasR protein, because till date there is no molecular detail information. Therefore, in this work we analyzed the interactions of the quorum sensing molecule N-3-Oxododecanoyl Homoserine Lactone (3-O-C12-HSL) with LasR. A new binding site of the 3-O-C12-HSL with the beta turns in the short linker region was found.

Introduction

<em>Pseudomonas aeruginosa</em> (<em>P. aeruginosa</em>) is a Gram-negative, monoflagellated, obligatory aerobic bacteria. The central problem is that it infects patients who are immunologically compromised, like those suffering from AIDS, cystic fibrosis, cancer, etc. There is an urgent requirement for novel antibacterial treatments to counter <em>P. Aeruginosa</em>. In this work, an attempt has been made to analyze molecular properties of the LasR protein as well as the mode of its interactions with 3-O-C12-HSL. We performed docking and molecular dynamics (MD) simulations of the LasR protein of <em>P. aeruginosa</em> with the 3-O-C12-HSL ligand. We assessed the conformational changes of the interaction of the 3-O-C12-HSL with LasR.
Methods

The protein LasR is a 239 residue long protein. The crystal structure of the N-terminal Al ligand binding domain (LBD) of LasR protein contains residues 1 to 169. However, in order to find the mode of interactions between the LasR protein and 3-O-C12-HSL, the structure of the entire LasR protein was needed. The homology modeling of LasR protein was done using the HHPred web server\(^1\). 3-O-C12-HSL parameters was generated using acpype\(^2\). Blind docking of 3-O-C12-HSL with LasR monomer was performed using Autodock Vina\(^3\). Centroid conformations from molecular docking were used as starting points for molecular dynamics (MD) simulations. MD simulations were performed using Gromacs\(^4\). PCA and cluster analysis were performed on docking conformations as well as MD trajectories\(^5\). The accuracy of the different clustering rounds was assessed with Davies–Bouldin Index (DBI), Silhouette Score, Dunn Index and Calinski-Harabasz metrics.

Results & Conclusions

In order to find the mode of interactions between the 3-O-C12 HSL and LasR protein, the structure of the entire LasR protein was reconstructed. We performed a simulation run of 100 ns using a standard MD protocol for the assessment of conformational changes of the reconstructed LasR monomer. Molecular docking and dynamics show that 3-O-C12 HSL can bind to multiple sites, not just LBD. So far this is the first report that shows that the 3-O-C12-HSL can interact as well with the beta turns in the the short linker region(SLR) of LasR (Figure 1).
This raises the question what would be the roles of different binding sites, a two step activation or a resistance mechanism. This study may therefore be used to develop new therapeutics to prevent the infections caused by \emph{P. aeruginosa} by targeting both the LBD as well as the beta turns in SLR and inhibit activation of genes.

**References**

51 - PhyrePower- Protein fold recognition by contact threading

Ioannis Filippis¹, Hongjiang Miao¹, Lawre¹, Michael JE Sternberg¹, Lawrence A Kelley¹

¹Structural Bioinformatics Group Department of Life Sciences Imperial College London South Kensington Campus London SW7 2AZ

Abstract

PhyrePower is a novel approach for predicting protein structure when there is no recognisable sequence similarity to a known structure. First protein contacts are predicted from multiply aligned sequences and then matched against a database of observed contact maps from the protein data bank. We considered 151 proteins whose structure could not be predicted by hidden Markov models searches against a structural database. PhyrePower identified at rank 1 a correct fold for 64 out of the 151 proteins (43%) and generated a good model for 34 out of these 151 proteins (23%).

Introduction

There are two main approaches to predict protein structure: template based identification by (multiple) sequence alignment and template free. A recent advance in template free is the use of predicted residue-residue contacts from which the model is constructed (e.g. EVFold [1] and MetaPSICOV [2]). In PhyrePower the two approaches are integrated.

Methods

Figure 1 overviews the method. First, a multiple sequence alignment is generated for the query. Next MetaPSICOV was used to predict a list of contacts ranked by their confidence scores. A contact is defined such that the Cbeta-Cbeta (Calpha for Gly) between two residues is ≤ 8Å. The template library was constructed from the SCOPe 2.04 domain database (all entries < 95% sequence identity). Contact maps were calculated for proteins in the template database. Contact map overlap (CMO) is defined as the number of common contacts between two maps divided by the number of contacts in the query. Each contact map is eigen-decomposed into the 7 principle eigenvectors. The method of DiLena et al [3] using dynamic programming was used to compare the query and a template contact map. The top 10 matched templates are used
to generate a set of alignments from different gap penalties. The alignment with the highest CMO is input into MODELER to yield the PhyrePower predicted structure.

The main application of PhyrePower will be when state-of-the-art sequence-based template methods fail. We reproduce this by establishing a benchmark data set of queries where (i) the query has < 20% HHSearch probability of matching any template domain, (ii) the query has a match in the template library with a TM score > 0.5 so a “good” model can be predicted, (iii) one representative is used per SCOP fold. This yielded 151 representative proteins whose structure could not be predicted by HHSearch fold recognition but for which there is a template in the library. PhyrePower was compared to constructing a model from the contacts using TINKER [4].

Results

At rank 1, PhyrePower identified a correct fold for 64 out of the 151 proteins (43%) and a good model (TM>0.5) for 34 out of the 151 queries (23%). PhyrePower yields prediction for all structural classes. The corresponding figures for TINKER are 55 correct folds (36%) and 32 good models (22%). Importantly, for correct folds there were only 31 queries in common between PhyrePower and TINKER and for good models only 16 queries in common. Thus PhyrePower predicts many folds and good models which were not identifiable by contact map prediction followed by TINKER. We have explored a strategy to combine PhyrePower and TINKER predictions and obtain a single prediction. This approach yielded correct folds for 78 queries and good models for 44 queries (29%). PhyrePower is available (for non-commercial use) as a Docker image

https://hub.docker.com/r/filippis/phyrepower-docker/
Figure 1 - Overview of PhyrePower

References

52 - Investigating the molecular determinants of ebolavirus pathogenicity

Mark Wass¹, Morena Pappalardo¹, Miguel Julia¹, Ian Reddin¹, Diego Cantoni¹, Francesca Collu², James MacPherson², Jeremy Rossman¹, Franca Fraternali², Martin Michaelis¹

¹University of Kent, ²King's College London

Abstract

The West Africa Ebola virus outbreak killed thousands of people. Using sequencing data combined with detailed structural analysis and experimental data, we compare Ebolavirus genomes to identify potential molecular determinants of Ebolavirus pathogenicity. We identify specificity determining positions (SDPs) that may act as molecular determinants of pathogenicity. Of 189 SDPs protein-structural analysis revealed eight that were likely to alter protein structure or function. SDPs present in VP24 are likely to impair binding to human karyopherin alpha proteins and prevent inhibition of interferon signaling in response to infection. Secondly structural analysis of the mutations present in Ebola during rodent adaptation experiments suggested that fewer than five mutations are required to introduce pathogenicity in a new host species. Mutations in VP24 are critical to adaptation. As only a few mutations are needed for adaptation and only a few SDPs distinguish Reston virus VP24 from other Ebolaviruses, it is possible that human pathogenic Reston viruses may emerge.

Introduction

The recent Ebola virus outbreak in West Africa demonstrated the ability for Ebola to be deadly on a large scale. Our interest is in identifying the molecular determinants of Ebolavirus pathogenicity. We utilised knowledge that of the five Ebolavirus species, only one of them, Reston, is not pathogenic in humans. We hypothesise that the differences between the genomes of Reston virus and the other Ebolavirus species must explain the difference in human pathogenicity. We combine protein structural analysis including molecular dynamics to investigate the differences between the two groups of Ebolaviruses.

Ebola is not pathogenic in rodent species. Studies have used serial passaging of the virus in rodents, the virus adapts to the rodent host and induces pathogenicity. We use
protein structural analysis to investigate how these mutations make the virus pathogenic and compare to the SDPs identified in our initial study.

Results & Conclusions

Our comparison of Reston viruses and human pathogenic Ebolaviruses identified 189 SDPs [1]. Structural analysis identified eight SDPs likely to affect protein function and the potential to alter pathogenicity between Reston and the other species. Four SDPs in VP24, three in the interface with the human protein karyopherin alpha5 (KPNA5) and one removes a hydrogen bond in VP24. VP24 is multifunctional including antagonism of the host interferon response. VP24 inhibits interferon signalling by binding both karyopherin proteins and STAT1, thus preventing STAT1 nuclear localisation and therefore blocking activation of the interferon response.

Analysis of the SDPs in the binding site suggested they were likely to affect binding with KPNA5. Molecular dynamics simulations of the VP24-KPNA5 complex suggest that the interaction between VP24 and KPNA5 is less stable [2]. This diminished interaction would reduce the ability of Reston viruses to prevent interferon signalling and could explain the lack of human pathogenicity in Reston viruses.

Analysis of the mutations that occur during rodent adaptation revealed multiple VP24 mutations occurred in similar locations to the SDPs (Figure 1); either in the interface or removing hydrogen bonds [3]. Our analyses of Ebolavirus mutations suggest that VP24 is key to determining host pathogenicity and that very few mutations are required to alter pathogenicity.
FIGURE 1. Mutations in VP24. The complex of Ebola virus VP24 (grey) and human KPNA5 (cyan) is shown, SDPs (red) and adaptation mutations (blue).

References


53 - Sphinx: Merging Knowledge-Based and Ab Initio Approaches to Improve Protein Loop Prediction

Claire Marks¹, Jaroslaw Nowak¹, Stefan Klostermann², Guy Georges², James Dunbar², Jiye Shi³, Sebastian Kelm³, Charlotte M Deane¹

¹University of Oxford, ²Roche Innovation Center Munich, ³UCB Pharma

Abstract

Loops are often important for protein function, but are difficult to model accurately due to their structural diversity. Currently, most loop structure prediction algorithms belong to one of two classes: knowledge-based and ab initio. Here we describe a novel algorithm, Sphinx, which combines aspects of these two approaches, producing high-accuracy predictions and decoy sets enriched with near-native conformations.

Introduction

Loops often play a vital role in protein function, and knowledge of their structures is important to understand a protein’s properties. However, due to the variability in their structures and sequences between homologues, they are usually the least accurate regions of a protein model.

The majority of existing loop structure prediction algorithms are of two distinct types: knowledge-based, where candidate conformations (decoys) are selected from databases of known protein fragments; and ab initio, where decoys are generated computationally from scratch. Each approach has its advantages and disadvantages. Knowledge-based algorithms are normally fast, and can provide high-accuracy predictions when the target loop structure is similar to one already observed, but fail when the target loop has a novel conformation. Further, they are not able to use fragments that are a different length to the target loop, even though it has been shown that loops of different lengths can adopt similar conformations. Ab initio methods are able to generate novel shapes, but must explore a very large conformational space, making them computationally expensive.

Methods
We have developed a novel algorithm, Sphinx\textsuperscript{2}, which is a hybrid of knowledge-based and \textit{ab initio} approaches. Sphinx is able to use the extra structural information contained within fragments with a different number of residues to the target, by copying structural information (dihedral angles etc.) from a fragment shorter than the target loop according to a sequence alignment, and completing the structure using \textit{ab initio} techniques.

\textbf{FIGURE 1.} An example of a prediction produced by Sphinx. The target loop is eight residues in length (residues 115-122 of PDB entry 2z9wA, left). FREAD, a knowledge-based method, failed to find any suitable length-matched fragments, and hence produced no prediction. However, there is a protein fragment that is structurally similar to part of the loop, but is two residues shorter (residues 65-70 of 4n05A, centre). Sphinx, unlike other loop modelling algorithms, is able to use this fragment, and produces an accurate prediction with an RMSD of 0.76 Å.

\textbf{Results & Conclusions}

Sphinx is able to provide high-accuracy predictions, often with an RMSD below 1 Å, achieving comparable results to the knowledge-based method on which it is based (FREAD\textsuperscript{3,4}), and considerably outperforming the \textit{ab initio} component. The generated decoy sets are enriched with near-native conformations relative to its base \textit{ab initio} algorithm, and it is able to produce a prediction in every case, unlike some knowledge-based methods which fail when no suitable length-matched fragment can be found. Sphinx also achieves comparable results to some leading algorithms, without prohibitive execution times.

\textbf{References}


Short linear motifs (SLiMs) are rapidly evolving stretches of amino acids that mediate transient low affinity protein-protein interactions. It is unclear though why certain interactions are highly specific than others. Calcineurin (Cn), a critical phosphatase targeted by immunosuppressant drugs, specifically recognizes a SLIM called PxIxIT in its interacting partners. Although different binding strengths of Cn-PxIxIT interactions are pivotal to the dynamics of Cn signaling, specificity landscape of these interactions has not been investigated systematically. We are combining powerful computational and experimental techniques to define and systematically characterize specificity determinants in Cn-PxIxIT interactions. Using flexible backbone modelling on the available structures of Cn-PxIxIT complexes, we first qualitatively predict residues tolerated at each of the positions in PxIxIT motifs. Next, starting from a previously known high affinity PxIxIT – PVIVIT – in complex with Cn, we compute the difference between theoretical free energies of wild type and single amino acid mutant peptides to quantitate the effect of tolerated and non-tolerated mutations on binding strengths. In parallel, we are developing a high throughput technology platform to chemically synthesize the modelled peptides on spectrally encoded beads and estimate with high specificity, the binding strengths of individual Cn-peptide interactions. Strong correlations between theoretical predictions and experimentally measured binding strengths have identified two variants of PVIVIT that further improve binding and also, a repertoire of amino acids that reduce binding. This positive and negative information is critical in developing in silico approaches for proteome wide prediction of PxIxIT motifs and the discovery of new Cn targets. Extending this approach to other protein-peptide interactions will potentially transform drug discovery efforts. Strong peptide binders can be further modified to design candidate inhibitors of protein interaction networks.
Abstract & Introduction

A key concept in Template-Based Modeling (TBM) is the high correlation between sequence and structural divergence. The main practical consequence of this correlation is that homologous proteins that are similar at the sequence level will also be similar at the structural level allowing the selection of a proper template for a target sequence. Pioneering work by Chothia and Lesk[1] found a non-linear and well correlated relationship between sequence and structural divergence. However, a given protein sequence could exists in different structures (conformers) where their structural differences describe their conformational diversity (CD). In this work, we explored the impact that CD has on the relationship between structural and sequence divergence.

Methods

CoDNaS database[2] was used to recruit proteins exhibiting conformational diversity. Maximum conformational diversity for each protein is the maximum C-alpha RMSD derived from all conformers pairwise comparisons. Using this set, we ran BLASTClust to obtain all available clusters at 30% of local sequence identity. The final dataset contains 2024 different protein chains with a total of 37755 conformers. These proteins are grouped in 524 families.

To estimate the structural divergence for each homologous protein pairs in a cluster, we calculated the C-alpha RMSD for all possible pairs of conformers belonging to the proteins being compared. Additionally, we calculated the percent of sequence identity for each homologous protein pairs using a global sequence alignment. The total comparisons among all vs all conformers for each homologous protein pairs and structures of the same protein give an amount of ~3.5 millions of pair.

Results & Conclusions

We found that the use of a highly redundant sequence dataset (that is, considering the CD) blurs the well-established relationship between sequence and structure divergence more than shown in previous studies (Fig. 1). It is also evident from Figure 1 that the
extent of conformational diversity can be as high as the maximum structural divergence among families reached by accumulation of nonsynonymous substitutions. Also, the presence of CD impairs the well-established correlation between sequence and structural divergence, which is more complex than previously suggested due to the existence of different structures (CD) for the same sequence. However, we found that this noise can be resolved using a priori information from the structure-function relationship. We showed that protein families with low CD, which we called “rigid”[3], show a well-correlated relationship between sequence and structural divergence (Spearman’s rank correlation rho of -0.83), which is severely reduced in proteins with larger CD (Spearman’s rho = -0.51). This lack of correlation could impair template-based modeling (TBM) results in highly dynamical proteins due to the uncertainty to select a proper target structure. Finally, as proteins with disordered regions show higher extensions of CD, we also found that the presence of order/disorder can provide useful beforehand information for better template selection and TBM performance.

**Figure 1.** Maximum RMSD (MSD and CD) versus sequence percent identity. Points refer to the maximum RMSD obtained from an “all vs all” comparison between structures from two homologous proteins (MSD), or from the same protein (CD). (a) Green dots: comparisons between homologous protein pairs (Spearman’s rho = -0.52). Red dots: comparison between conformers of the same protein. (b) Distributions of the maximum RMSD between two homologous proteins (green) and between conformers of the same protein (red).

**References**


58 - Inferring protein phylogeny by modelling the evolution of secondary structure

Jhih-Siang Lai¹, Bostjan Kobe¹, Mikael Boden¹

¹School of Chemistry and Molecular Biosciences, The University of Queensland, Australia

Introduction

Ancestral sequence reconstruction (ASR) has had recent success in decoding the origins and the determinants of complex protein functions. However, attempts to reconstruct extremely ancient proteins and phylogenetic analyses of remote homologues must deal with the sequence diversity that results from extended periods of evolutionary change.

In the last 20 years, the number of protein structures in the PDB has increased 20-fold. Seizing wealth of protein structures, we propose to cast structural change over evolutionary time as substitutions between secondary structure (SS) states. Dayhoff’s point accepted mutation (PAM) matrix describes evolutionary changes for amino acids (AAs) and assigns a probability to evolutionary events. This evolutionary model is based on differences between discrete states observed in modern proteins and those hypothesized in their immediate ancestors. The approach naturally extends to structural traits, assuming access to numerous high-resolution homologous protein structures.

We present the first protein structural evolutionary model in SS terms, and we test its capability for inferring phylogeny and ancestral SS from sequence-diverse but structurally important protein families.

Methods and Results

For all protein structures, DSSP was used to assign each residue one of 7 SS states, i.e., beta bridge, strand, helix-3, alpha helix, helix-5, bend, and turn. We clustered proteins at 85% AA sequence identity, resulting in a dataset with 75 clusters and 592 proteins. For each cluster, we built a maximum parsimony tree based on the multiple sequence alignment (MSA). Following Dayhoff’s approach, we counted putative
evolutionary SS changes to calculate a SS 1 PAM, excluding sites with unknown SS states (Figure 1A).

To test how the structural evolutionary model to complement analyses by AAs, we implemented a phylogenetic tree inference algorithm. Our method is based on maximum likelihood and takes a multiple structural alignment to infer all branches of a tree.

Toll/interleukin-1 receptor (TIR) domains are present across 3 kingdoms; they are structurally similar but extremely sequence-diverse (identity <20%). By performing analyses using both a SS-based and a AA-based evolutionary model, we recovered biological properties of TIR domains from 28 structures. For instance, all TIR domains of human adaptors diverge from the same origin (Figure 1B); with the AA model they are mixed up with receptors. TRAM and TRIF are separated to have evolved interaction with NF-kB and IRF3 signalling pathways. Bacterial domains are well demarcated. The structural evolutionary model is clearly able to extract phylogenetic relationships in sequence-diverse families.

To show our model can capture structural evolutionary changes, we developed an algorithm to infer ancestral proteins’ SS. We used published data from three independent ASR studies for a protein kinase, a polar amino acid-binding protein, and a DNA-binding domain. We made inferences by maximizing their joint probability with our model, given the published phylogenetic trees and multiple secondary structure alignments (MSSAs), populated with modern protein structures. The MSSAs were created by substituting AAs in the published MSA for SS.

We compared the crystallized ancestral structures from the three studies above, with our ancestral SS inferences; we used homology modeling and AA-based SS predictors to quantify how close we can get to the ancestral structures. SWISS-MODEL built 3D structure by assigning each modern structure as template. PROF, JPred, and NetSurfP predicted SS on the published ancestral AA sequences. Our inferences outperform homology modeling for 5 of 7 ancestors by >15% accuracy and give comparable performances for the other two (Figure 1C). Our inferences outperform all AA-based predictions (Figure 1D). Hence, the structural evolutionary model extracts information not available from the modern structures or the ancestral AA sequences alone.
Figure 1. (A) 1 PAM for secondary structure. The values are multiplied by 10,000. The 1 PAM describes the evolutionary changes in terms of transition probability at ≤0.01. The 1 PAM is interpreted columnwise, e.g., 1 PAM shows that B has a probability 0.0177 change to E. (B) Structural phylogenetic tree of TIR domains. (C) Prediction accuracy for ancestral SS inference and homology modeling. We averaged the accuracy of template-based predictions for each ancestral protein structure. (D) Prediction accuracy for ancestral SS inference and sequence-based SS methods. Seven-class accuracy is defined by DSSP’s annotation. Three-class accuracy is defined by beta sheet (B or E), helix (G, H, or I), and other (S or T).
60 - New Insights into statistical potentials for describing protein binding affinity and aggregation properties

Fabrizio Pucci¹, Qingzhen Hou¹, Jean-Marc Kwasigroch¹, Marianne Rooman¹

¹Université Libre de Bruxelles

Abstract

Statistical potentials are a widely used tool in structural bioinformatics since they represent an accurate and efficient method of investigation that can be used at structuromic scales. Here we review their peculiar characteristics that allow probing solubility and binding affinity characteristics of amino acid interactions, and present some recent developments for protein design.

Introduction

Since their introduction about 30 years ago¹,², statistical potentials represent one of the powerful methods in structural bioinformatics to analyze protein biophysical characteristics. Despite the approximations on which their construction is based, they are widely used in a lot of applications for their accuracy and their computational efficiency. These two characteristics make them a perfect instrument especially for large-scale, structuromic, investigations. Here we present some recent applications that we developed using these potentials of mean force for the analysis of protein solubility and protein-protein interactions.

Methods

The statistical potentials that we have built are mean force potentials derived from datasets of known protein 3D structures and are based on a coarse-grained representation, in which the side chains are simplified by average side chain centroids. Denoting by \( s \) a sequence element such as a single amino acid or an amino acid pair, and by \( c \) a structure element such as an interresidue distance, a solvent accessibility range or a backbone torsion angle domain, the energetic contribution associated to the configuration \((c, s)\) is obtained from the inverse Boltzmann law:
$$\Delta W(c,s) = - k T \log[P(c,s)/(P(c)P(s))]$$

where $P(c,s)$, $P(c)$ and $P(s)$ are the probabilities to observe $(c,s)$, $(c)$ or $(s)$, $k$ is the Boltzmann constant and $T$ the absolute temperature. These probabilities are approximated in terms of the number of occurrences of these sequence and structure elements in a dataset of experimentally resolved protein structures. By exploiting the bias of the potentials towards the dataset\textsuperscript{3,4} from which they are derived, we defined new interface- and solubility-dependent potentials that describe the binding affinity and aggregation properties of proteins.

**Results & Conclusions**

First we present results derived from the analysis of our interface- and solubility-dependent statistical potentials regarding the relevance of some specific interactions such as salt bridges and cation-pi interactions, in the modulation of protein-protein affinity and protein solubility. Then we discuss how these potentials can be fruitfully applied for protein design. In particular, we present an optimized version of our bioinformatics tool for the prediction of binding affinity changes upon mutations (called BeatMuSiC\textsuperscript{5}), which incorporates these new potentials and is one of the most precise and fastest tools in the literature. Finally we show preliminary results on its application to the analysis of the interactome’s stability.

**References**

61 - Using Local States to Drive the Sampling of Global Conformations in Proteins

Alessandro Pandini¹, Arianna Fornili²,³

¹Brunel University London, ²Queen Mary University of London, ³The Thomas Young Centre for Theory and Simulation of Materials, London

Abstract

We present a novel approach to drive the sampling of conformational transitions in Molecular Dynamics (MD) simulations using libraries of local states in the form of Structural Alphabets (SA). The reported findings indicate that molecular simulations and knowledge-based fragment libraries can be effectively combined to enhance the exploration of the conformational space of proteins. The proposed strategy has a wide range of applications, ranging from protein model refinement to protein folding and design.

Introduction

Conformational changes in proteins are often associated with protein-protein interactions, ligand binding, and post-translational modifications. The structural and energetic characterization of conformational transitions is therefore of central interest in understanding protein function. However, these transitions often occur at timescales inaccessible to unbiased MD simulations, consequently different approaches have been developed to accelerate their sampling¹. Here we investigate how knowledge of experimental backbone conformations preferentially adopted by protein fragments, as contained in pre-calculated SA libraries², can be used to explore the landscape of global protein conformations in MD simulations.

Methods

SAs were successfully used to analyze trajectories from MD simulations³,⁴. Here we define a novel SA-based Collective Variable (CVSA)⁵ to bias the sampling of backbone conformations of protein fragments towards recurring local states found in experimental structures. Examples of use in Metadynamics and Steered MD for two simulation challenges are presented: folding of small proteins and structural refinement of protein models.
Results & Conclusions

We find that: a) Enhancing the sampling of native local states allows recovery of global folded states when the local states are encoded by strings of SA letters derived from the native structures. b) Global folded states are still recovered when the information on the native local states is reduced by using a low-resolution SA, where the original letters are clustered into macrostates. The macrostates provide the approximate shape of the fragments, while sampling with the atomistic force field allows the structure to adopt the native conformation of the specific amino acid sequence. c) SA strings derived from collections of experimental structural motifs can be used to sample alternative conformations of pre-selected regions. We are currently extending our approach by combining the CVSA with contact prediction from residue coevolution methods.

**FIGURE 1.** “Schematic description of the CVSA and its applications” by Pandini & Formili is licensed under CC-BY 4.0. (A) The plot of the switching function used in the definition of the CVSA is reported, showing the dependence of the function on the Ca RMSD (ρ) of a fragment (gray to dark green licorice) from a reference SA state (green). Using the CVSA, the sampling of local states can be biased for all the protein fragments in folding simulations (B) or for fragments in selected regions during structural refinement (C).
References


Enteroviruses, a genus of the Picornaviridae family, cause many human diseases. Currently there are no antivirals against enterovirus infections. Capsid expansion is critical for the release of RNA into the host cell. Although this process hints at possible drug targets, it remains poorly understood. As a model, we investigated capsid expansion of Enterovirus 71 by Normal Mode Analysis and Perturbation Response Scanning (PRS). We also conducted a bioinformatic screen of all available sequences of enterovirus capsid proteins. We identified the dominant motions and conserved hotspots that may function in capsid expansion. We propose that expansion may be altered by drugs targeted at these regions. Our approach is computationally feasible and can be applied to other virus families.

Introduction

Enterovirus capsids are non-enveloped icosahedrons. The asymmetric unit is a protomer comprising of four subunits VP1-VP4. The capsid is assembled as 12 pentamers, each pentamer containing five protomers. The external subunits are VP1-VP3, while VP4 is an internal protein. The capsid has two-fold, three-fold and five-fold axes of symmetry. Upon binding to the host cell receptor, conformational changes within the capsid are triggered, resulting in an expanded intermediate capsid. This RNA-release intermediate is termed the A-particle. The receptor binding site is thought to be a canyon-like depression, surrounding the five-fold axis [1], while structural studies indicate pore formation at the two-fold axis [2]. We investigated the motions and residues responsible for the expansion between the full capsid and its A-particle.

Methods

The normal modes of various coarse grained complexes of the RNA-containing capsid of Enterovirus 71 were investigated using the Anisotropic Network Model (ANM) [3]. This
included an individual protomer, a pentamer and coarse grained viral capsids. Also, to inspect the modes associated with pore formation at the two-fold axis, we analysed a sub-complex of three interacting pentamers. To reduce computational expense we also applied ANM to further coarse grain complexes comprising of a subset of the Cβs equally distributed in three-dimensional space. Key residues associated with expansion were identified by PRS [4] and mapped to conserved motifs following a comprehensive sequence analysis [5].

Results and Conclusions

We identified a common mode in the capsid, interface and pentamer structures with high overlap to capsid expansion. Visualisation of eigenvectors indicated outward expansion at the five-fold axis, with collapsed regions at the pentameric interface (Fig 1). PRS identified a channel of conserved residues at the five-fold axis as a hotspot for pentamer expansion. Common modes were obtained regardless of the degree of coarse graining; thus we present a computationally feasible method to analyse the motions of viral capsids. Our results advance the understanding of viral uncoating and suggest the five-fold axis as a primary region for motions associated with RNA release.
Figure 1. Motions and hotspots associated with RNA-release in complexes of the Enterovirus 71 capsid. Red) Native full capsid. Blue) RNA-release intermediate. Lightblue) Hotspots identified by PRS.

References


63 - The determination of force field parameters of the conserved copper coordinating active site of AA9 proteins

Vuyani Moses¹, Ozlem Tastan Bishop¹, Kevin Lobb¹

¹Rhodes University

Abstract

The Auxiliary Activity family 9 (AA9) proteins are crucial for the early stages of cellulose degradation. The presence of a Cu²⁺ atom in the active site of AA9 proteins makes these enzymes difficult to study using Molecular Dynamics (MD) simulations. As a result force field parameters for the Cu²⁺ Type 1 active site have been evaluated and validated on all three AA9 types. Once complete, the MD trajectories for all three AA9 types were assessed and Type specific interactions were identified.

Introduction

AA9 proteins are metal coordinating enzymes that have been shown to have a positive effect on cellulose degradation when present with traditional cellulases. The presence of a Cu²⁺ metal in the AA9 active site grants these proteins monooxygenase activity. The exact mechanism in which AA9 proteins degrade cellulose remains elusive however, it is known that three cleavage products are formed when interacting with cellulose. Type 1 AA9 proteins produce a C1 cleavage product, Type 2 AA9 proteins produce a C4 cleavage product and Type 3 AA9 produce both C1 and C4 cleavage products [1]. Our previous work has identified type specific sequence and structural features that may play a role in regioselectivity of AA9 proteins [2]. As a result, force field parameters were generated for Type 1 AA9 proteins [3] and validated on all three AA9 types in order to study the interaction between AA9 proteins and cellulose.

Methods

Sequence alignments, motif analysis and phylogenetic analysis were used to identify AA9 type specific features of AA9 proteins. MD simulations were then used to study AA9 - substrate interaction. The force field parameters for copper coordinating bonds were calculated for the Type 1 AA9 active site. Using the new force field parameters, MD
Simulations were performed for all three types using a three layered β-cellulose crystal. MD simulations revealed type specific interactions between AA9 types and cellulose.

Results & Conclusions

PES scans at the semi empirical PM6 level of theory were used to generate the copper force field parameters. The parameters were then validated with MD simulations for all three types. Two Lennard-Jones parameter sets were used for the analysis. This resulted in two MD simulations for each AA9 type which were referred to as biased and unbiased MD experiments. The interaction between the Type 1 AA9 protein and cellulose was assessed and is shown in Figure 1.

AA9 proteins were found to have motion relative to the cellulose substrate. Copper binding to the free hydroxyl on the cellulose was observed for Type 1 proteins. The use of a biased Lennard-Jones parameters was found to result in a rapid binding to cellulose.

References


β-sheets peptides (BPs) have been used to stabilize integral membrane proteins (IMPs) by sequestering the hydrophobic surfaces [1]. We employed a systematic computational approach using molecular dynamics (MD) simulations to design β-sheets and study their formation in aqueous solutions. The MD data provides new insights into the structure and dynamics of β-sheets, which is possibly relevant to the stabilizing effects of β-sheets on IMPs, as well as a starting point for modeling β-sheet/IMP complexes.

Introduction

IMPs play crucial roles in all cells. However, functional and structural studies of IMPs are hindered by their hydrophobic nature and the fact that they are generally unstable following extraction from the membrane environment. Recently, BPs were used to maintain IMPs stable [1]. These BPs are 8-amino-acid peptides with alternating polar and apolar residues with an octyl side chain at each end. The major incentive to explore the β-sheets and its derivatization with functionalized groups is that they may enable stoichiometric and oriented crosslinking of IMP’s with the solid substrate, for example by designing the functionalized parallel β-sheets. However, it is extremely time-consuming, and most often unrewarding, to systematically explore the effect of modifications to the chemical structure of β-sheets such as the length of the β-sheet, the distribution of hydrophilic and hydrophobic residues and the inter-strand hydrogen bond interactions experimentally. Here, we employed MD simulations to investigate the β-sheet formation of BP1 as well as two other modifications of BP1 namely, propargyl-BP1 and azido-BP1 by adding a propargyl/azido group at the N-term of BP1. The latter two functionalized BPs are designed in a way to not only stabilize the membrane protein but also provide a means to covalently immobilize the IMPs on a solid substrate. We are planning to use the AqpZ, the water channel from *E. coli* as a model protein in future to study β-sheet/IMP complexes. The structure and function of AqpZ has been well studied through MD
simulation, suggesting that we can use MD to address the effect of the β-sheets on AqpZ behaviour.

Methods

We designed and built the β-sheets by docking peptides. We employed MD to study the stability and properties of the β-sheets using AMBER code [2]. We introduced different non-standard residues to the AMBER forcefield using PyRED software [3]. Beta conformational states were defined by Φ and ψ torsion angles (−180°<Φ<−90° and 50°<ψ<180°). The cutoff for donor–acceptor distance was set to 3.5 Å as a criterion for the hydrogen bond and no angular restrictions were applied. Energy evaluation was conducted on β-sheets using the simulations with explicit solvent, energy terms were evaluated using the MM-PB(GB)SA [4] methods.

Results & Conclusions

We found that differences in functionalized groups and amino acid insertions in the sequence of peptide could lead to changes in the size and orientation of the β-sheet assemblies. It would be very interesting to take such criteria to systematically explore the effect of variations in chemical structure, so as to help the experimentalists to develop better β-sheets to stabilize IMPs and control their orientation on solid surfaces. A model of the BP1 β-sheet is shown in the Figure 1.
Figure 1. BP1 (a) peptide (b) β-sheet.

References

65 - Determining Allosteric Hot Spots in Hsp70 using Perturbation Response Scanning

David Penkler¹, Ozge Sensoy², Canan Atilgan⁴, Ozlem Tastan Bishop³

¹Rhodes University, ²Istanbul Medipol University, ³Rhodes University, ⁴Sabanci University

Abstract

The molecular chaperone Hsp70 is a highly conserved housekeeping protein, involved in the folding and maturation of various peptides. Central to Hsp70’s function is its ability to bind, and release peptides by interconverting between two distinct functional states via an allosteric mechanism. We combine molecular simulations with Perturbation Response Scanning (PRS) as a viable approach to probe the intricacies of Hsp70’s complex allosteric mechanism. We show that simultaneously bound ATP and peptide allosterically activate the complex and report several allosteric hot spots that favour inter-state conversion when externally perturbed. Our PRS data closely fits the current opinion of Hsp70’s functional cycle and we report residues that may be involved in the allosteric regulation of Hsp70.

Introduction

Hsp70s are able to rescue misfolded, partially denatured or aggregated proteins by mediating their unfolding and refolding via an ATP driven allosteric mechanism¹. Hsp70’s client array includes known oncogenic and disease related peptides, making it an attractive drug target. Clients are bound at the C-terminal substrate binding domain (SBD), while nucleotide is bound at the N-terminal ATPase domain (NBD). Allosteric communication between these domains is crucial to the binding and release mechanism, allowing the protein to interconvert between an ADP and substrate bound closed conformation², and an ATP bound open conformation³. We employ PRS⁴,⁵ on several ligand bound configurations of DnaK to investigate their respective allosteric potential and shed light on the intricate regulation of this complex machine.

Methods

Structural configurations of DnaK were constructed based on experimental structures (PDB 2KHO and 4B9Q) by adjusting the bound state of the NBD and SBD. Each complex was simulated using molecular dynamics and the resulting trajectory was used in a series of PRS experiments. PRS is a computational technique that measures the response of all residues by scanning each residue with force perturbations. Correlating
this response to a known target conformation, PRS may be used to determine the allostERIC influence each residue has on all other residues.

**Figure 1. Graphical representation of PRS data**

**Results & Conclusions**

In this study, we analyse the propensity of each configuration to interconvert to its opposite state (6 opening and 6 closing transitions). For both transition types, PRS data closely fits the current opinion of Hsp70 conformational cycle. Bound ADP or an apo NBD, regardless of peptide client, appears to favour the closed state. While ATP without peptide stabilises the open state. Interestingly, combining ATP and client substrate appears to allosterically activate both conformational states, potentially priming them for inter-state conversion via an intermediate structure. PRS employed on these complexes accentuates several residues known to be involved in allostERIC signal transduction as well as several residues unique to this study, including subdomain interface residues in the open state and several residues in the substrate entrapping α-helix of the closed state (see Figure 1). We hypothesise that these residues may play a role in regulating allostERIC control and that their perturbation may occur naturally via co-chaperone binding.

Overall, our study demonstrates how PRS in conjunction with MD can be used to
quantitatively assess the allosteric potential of a large multi-domain protein and identify individual residues that may be implicated in allosteric signal transduction.

References
1. Kityk, R; Vogel, M; Schlecht, R; Bukau, B; Mayer, M. Nat. Commun. 2015, 6, 8308.
5. Atilgan, C; Gerek, Z; Ozkan, S; Atilgan, A. Biophys. J. 2010, 99, 933–943
Prediction of differentially expressed genes from PCOS women by analyzing comorbidities associated with the disease

Ettahi Khaoula², Catherine Jennita³, Ashvini Desai¹

¹VIT University, ²Informatics and Data Analysis, ENSAT School, Abdelmalek Essadi University, ³Stella Maris College 17, Cathedral Rd, Chokkalingam Nagar, Teynampet, Chennai, Tamil Nadu 600086

Polycystic ovary syndrome (PCOS) is one of the most common metabolic and reproductive disorders among women of reproductive age and is linked to various associated disease or disorders such as obesity, insulin resistance, Type II diabetes mellitus, cardiovascular diseases, infertility, cancer, psychological well being, etc. current study was aim to explore the relation between obesity and PCOS by identifying the novel biomarkers for PCOS, specifically correlated with obese women. The datasets were normalized by RMA (Robust Multichip Average) Algorithm for Differentially expressed genes (DEGs) selection between disease and control patients using t-test and fold-change criteria simultaneously. EnrichR and DAVID online tools were employed for genes annotation and pathway enrichment analysis. 28 DEGs were selected with the relation between obesity and psychiatric disorders co-morbid with PCOS. Also, lung was the most affected organ by the DEGs in the gene enrichment studies followed by ovary and urinary tract, central nervous system, lymphoid and hematopoietic tissue in obese women with PCOS. 9 out of 28 DEGs were associated with pathways that are associated with obesity. 4 genes were enriched with positive regulation of Chemotaxis, a biological process intensely affected in obese women with PCOS. Thus, alteration in the expression of several genes and associated pathways clearly depicts a relation between obesity and PCOS with other risk factors involved with the disease.
70 - Protein Structures and their features in UniProtKB

UniProt Consortium$^{1,2,3}$, Nidhi Tyagi$^1$

$^1$EMBL-European Bioinformatics Institute, Cambridge, UK, $^2$Swiss Institute of Bioinformatics, Geneva, Switzerland, $^3$Protein Information Resource, Georgetown University, Washington DC & University of Delaware, USA

Abstract

Annotation of proteins from structure-based analyses is an integral component of the UniProt Knowledgebase (UniProtKB). UniProt works closely with the Protein Databank in Europe (PDBe) to map 3D structural entries (~100,000) to the corresponding UniProtKB accessions accurately and to import relevant data from the protein structures.

Introduction

With increasing submission of protein 3D structures to PDB every week, there is a need for correct annotation of these biomolecules. UniProtKB provides a comprehensive and thoroughly annotated protein resource to the scientific community. The reviewed UniProtKB/SwissProt section provides expertly curated annotation for each protein entry while the unreviewed UniProtKB/TrEMBL section contains automatically generated data. UniProt ensures accurate mapping of PDB structures to UniProtKB records which facilitates linking between the resources, and importing of data such as ligand-binding sites from PDB protein structures into UniProtKB.

Methods

PDB to UniProt mappings are carried out as a part of the SIFTS (Structure Integration with Function, Taxonomy and Sequence) project. SIFTS integrates relevant biological information for 3D structures from various databases such as IntEnz, Pfam, InterPro, CATH, SCOP and the Gene Ontology. The criteria to map a PDB entry to UniProtKB entry are as follows: (a) Two sequences should share high sequence identity (>90%) b) Mapping preference is given to UniProtKB entries from reference and complete proteomes c) Mapping is done at exact taxonomical level (strain level for lower organisms) d) Mapping is done to the longest protein sequence. A semi-automatic mapping pipeline has been developed which allows for automated mapping of many structures with manual intervention required for more complex cases. This accurate
mapping allows for the establishment of accurate cross-references between UniProtKB and PDB and for the subsequent import of small molecule data such as ligand-binding sites from PDBe.

Results & Conclusions

To date, UniProt has successfully completed the non-trivial and labour intensive exercise of cross-referencing ~350,000 polypeptide chains and 122,800 PDB entries to 40795 UniProtKB entries. Proteins with a solved 3D structure are priorities for expert curation and addition to UniProtKB/Swiss-Prot. During the curation process, use is of made of PDB data to enrich the records with information originating from the protein structure such as ligand-binding and active sites. In addition, data from PDB protein structures has been automatically imported into UniProtKB/TrEMBL, enriching the annotation of unreviewed records (16, 667 for May 2017 release). Protein structural information in UniProt thus serves as a vital dataset for various academic and biomedical research projects.

FIGURE 1. Cross referencing of PDB entry 3TIA in UniProtKB (entry Q194T1). Mannose binding site, glycosylation site, metal binding site and disulphide bond are annotated and visualised in the UniProtKB entry.

References


Introduction

Developing novel drugs in response to changing clinical requirements is a complex and costly process with uncertain outcomes. Computational modeling methods have increasingly been employed for hit identification and lead optimization. Studies estimate that each drug on average binds to at least six known protein targets and many other targets are yet to be discovered [1]. Predicting and minimizing the off-target interactions is important as these result in low efficacy, high toxicity and high failure rate of new drugs in clinical trials. We have developed a computational method that constructs structural signatures of small-molecule compounds that allows weakly conserved but functionally significant patterns to emerge and to be quantified.

Methods

After identifying the drug–target complex for each drug, we extracted the pocket that the drug binds to in the protein structure using the CASTp webserver [2]. We used sequence order-independent alignment to choose the pocket most similar to the bound pocket. The details of the construction of pocket ensembles can be found in [3].

Results & Conclusions

To capture the binding site features of a promiscuous drug, we use sequence order–independent structure alignment, hierarchical clustering and probabilistic sequence similarity to construct the PPE of the respective drug [2]. The PPE represents a unified set of individual pockets that potentially bind to several conformations of the drug. We have used the same approach to construct a negative PPE that represents a structural footprint of near native negative protein binding sites. Fig 1 shows the positive and negative PPE of sorafenib, which is approved for the treatment of advanced renal cell carcinoma and is on fast track for the treatment of advanced hepatocellular carcinoma. We were able to predict the interaction of six FDA approved kinase inhibitors (Sorafenib, Imatinib, Gefitinib, Dasatinib, Sunitinib, Pazopanib) with its known targets with a
sensitivity of 50% and specificity of 56%. This is significantly better than the precision (30%) and recall (27%) of a recently published method [4]. This methodology is general and can be broadly applied to small molecules for which a protein-small molecule bound structure has been solved.

Figure 1. The structural signatures of sorafenib, imatinib, dasatinib and gefitinib.

We validated our results for the top five novel targets for sorafenib using microscale thermophoresis experiments. Results suggest that two kinases (PKC eta and MAPKAPK2, Figure 2) have Kd values between 1-4 μM respectively, similar to known kinase targets for sorafenib. Similarly, we found that sorafenib interacts with two immune response related proteins with Kd values ranging between 300-600 μM. This sorafenib-immune response related protein interaction is also supported in literature. These novel interactions shed light on previously unexplained biology and side effects of sorafenib.

Figure 2. The predicted interaction of sorafenib with PKC eta and MAPKAPK2 proteins was experimentally verified through MST experiments. The affinities of PKC eta and
MAPKAPK2 with sorafenib was measured to be 1.09±0.36 μM and 3.68±0.14 μM respectively.

The challenge of identifying new drug–target pairs in silico has attracted significant interest from the computational community. Compared with our methodology, other existing methods including conventional docking or structure-based virtual screening, share one or more of the following limitations: (i) they are known to scale poorly with the size and complexity of drugs and drug binding sites and (ii) their algorithms do not appropriately account for the different conformations of both drug and binding site residues.

References

74 - LiteMol suite: A comprehensive platform for fast delivery and visualization of macromolecular structure data

David Sehnal¹,², Mandar Deshpande³, Radka Svobodova¹,², Saqib Mir³, Karel Berka⁴, Adam Midik¹,², Lukáš Pravda¹,², Sameer Velankar³, Jaroslav Koča¹,²

¹CEITEC - Central European Institute of Technology, Masaryk University, ²National Centre for Biomolecular Research, Faculty of Science, Masaryk University, ³European Bioinformatics Institute, ⁴RCPTM, Palacký University

Abstract

Visualization of the macromolecular structure data and its biological context is a critical step in their understanding. We have developed a complete solution that addresses the scientific challenge of providing an integrated view of macromolecular structure information and greatly outperforms currently available solutions.

Introduction

Recent advances in 3D structure determination techniques have facilitated the study of large macromolecular machines, resulting in a rapid increase in the number, size, and complexity of biomacromolecular structures available in the Protein Data Bank (PDB) and Electron Microscopy Data Bank (EMDB). In parallel, data resources (e.g., UniProt, Pfam) provide an important biological context for the PDB coordinate models.

Visualization of the macromolecular structure data and its biological context is a critical step in understanding and making effective use of this data. The ever-growing potential of use of web browsers and mobile devices necessitates a new generation of interactive and responsive visualization tools that can handle the increasing size and complexity of the available structures. These tools must enable fast delivery of the data and interactive 3D visualization. Furthermore, they have to be able to display additional information about the biological context and the underlying experimental data. Nevertheless, the currently available solutions for visualization (e.g., NGL, MolMil, JSmol) and data delivery (the MMTF format (1)) address this challenge only partially.
We have developed the LiteMol suite, an innovative solution consisting of a 3D molecular visualizer (LiteMol Viewer), data delivery services (CoordinateServer and DensityServer), and new data compression format (BinaryCIF) using standard mmCIF data format (2).

Methods

LiteMol Viewer can display 3D coordinate data, overlay them with additional annotations and readily displays experimental data. CoordinateServer and DensityServer provide on-demand access to relevant coordinate and experimental data available in the PDB and EMDB archives in required quality for specific use. For example, CoordinateServer can be requested to send only those atomic coordinates necessary for visualization of a cartoon model or only a ligand binding site. BinaryCIF is a compressed data format, which provides a compact way of storing and transferring the structural data.

Results & Conclusions
FIGURE 1. a) The impact of using CoordinateServer and BinaryCIF format on the size of transferred data for HIV-1 capsid (PDB ID 3J3Q). An interactive version at https://goo.gl/uduVJ4. b) The impact of CoordinateServer, DensityServer, and BinaryCIF on the size of transferred data for the Cryo-EM structure of the Zika Virus (PDB ID 5IRE; assembly 1). An interactive version is available at https://goo.gl/wNgBvW.

The LiteMol suite provides a complete solution to view 3D structure data interactively at all resolution scales, together with biological context in the form of annotations and with associated experimental evidence in the form of electron density or electric potential maps.

The LiteMol suite builds on the existing wwPDB standards, is up to orders of magnitude faster than the previous solutions, and enables near-instant delivery of model and experimental data, providing users with an interactive and responsive experience in all modern web browsers and mobile devices. The LiteMol suite is integrated into PDB in Europe (PDBe) and other life science web applications (e.g., SIB and CNRS services), and is freely available at www.litemol.org.

References

75 - Automated Realization of RNA Structure from Interaction Topology

Matthew Wicker¹, Di Chang¹, Russell Malmberg², Cory Momany³, Liming Cai¹

¹Dept. of Computer Science, Univ. of Georgia, USA, ²Dept. of Plant Biology, Univ. of Georgia, USA, ³Dept. of Pharmaceutical and Biomedical Sciences, Univ. of Georgia, USA

Background

The need for advanced methods of computational prediction of RNA structure cannot be overstated, given the critical role of RNA in cellular functions and the difficulty of experimental structure validation. Methods of outer-planar visualization, represented by NAVeiw [1], can illustrate canonical base pairs for predicted RNA secondary structure, but may not serve biologists well in functional segmentation and identification of higher order structural organization. Unfortunately, there has yet to be a direct way to render secondary structures into a three-dimensional model, which may consider additional, yet non-specific, tertiary nucleotide interactions [2]. Fragment assembly and molecular dynamic (MD) simulation deal with the problem, nevertheless in an inconvenient fashion. MD simulation consumes considerable amounts of time, and fragment assembly requires specific knowledge of tertiary interactions.

Methods and results

Our method allows for fast and accurate rendering of a meaningful 2D visualization that can be analyzed and directly translated into a full atomic 3D model. From an RNA sequence and a list of interactions across the nucleotide bases, sugars, and phosphates, we reduce the RNA sequence and its interactions to a graph by considering each nucleotide as a vertex and every interaction as an edge. Applying a biologically constrained version of the Kamada-Kawai algorithm [3] using the stress majorization method implemented in [4], we are able to realize the impacts of any tertiary interactions that exist in the interaction topology. Based on a geometric analysis of our improved outer-planar layout, we are able to regularly classify RNA pseudo-knots, junctions, and complex 3D motifs. This classification enables an accurate, monomer-by-monomer construction of an all-atoms 3D model. Comparisons of our 3D model to other methods
and real structures (rotated images) are given in Figures 1 and 2, respectively.

**Figure 1.** PDB ID 1L2X. Right to left: imposed tertiary interactions on NAView, our 2D drawing, Kinefold visualization [5].

**Figure 2.** PDB ID 1YMO. Top: Our structural embedding, real 3D structure. Bottom: Vfold [7] assembly, RNAComposer [8] assembly.

**Conclusions**

Our novel approach to RNA structure embedding in both two and three dimensions shows considerable improvements over existing methods, especially when information
about tertiary interactions is not complete, such as those given in the recent work based
on co-evolution analysis [2].

References


(2015). Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary


applications to software engineering. *Software Practice and Experience*, 30(11), 1203-
1233.

RNA/DNA folding path and structure prediction including pseudoknots and knots.

6. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K. J., Lukasiak, P., Bartol, N.,

Abstract
Antibodies are an important class of biopharmaceuticals. Currently, experimental antibody design is resource-intensive, and computational antibody design methods offer promise to facilitate this process. In particular, accurate structural modelling is essential. Here we describe our antibody-specific side chain predictor, PEARS, as an extension of our ABodyBuilder methodology. The two tools are designed to work with large volumes of antibody sequence data (e.g. next-generation sequence data) to pave the way for next-generation structure-based antibody design.

Introduction
Antibodies are proteins of the adaptive immune system. They are characterised by their high specificity and binding affinity, and can be designed to bind almost any target. However, experimentally designing a new therapeutic antibody is time-consuming and requires significant resources. Recent advances in next-generation sequencing (NGS) technologies and computational methods offer possibilities for tailoring antibodies for specific disease applications. Previously, we have shown that ABodyBuilder can build a model antibody structure for a paired antibody sequence within 30 seconds [1]. This is especially useful for NGS datasets [2], allowing users to map patterns from large volumes of antibody sequence data onto structural models. We have now added an antibody-specific side chain prediction step, PEARS, to not only enhance the accuracy of the models, but to also have a platform for generating single-mutant antibodies. PEARS harnesses the IMGT position-dependent distribution of rotamers to predict side chains. On model structures, PEARS is superior to other leading side chain predictors.

Methods
Our antibody-specific rotamer library was developed on a non-redundant set of 639 structures from SAbDab [3]. Unlike other rotamer libraries, we bin antibody rotamers according to their IMGT position, rather than their backbone dihedrals. We found that at specific positions, certain amino acids have a unimodal $\chi_1$ angle distribution; for example, glutamine residues in H44 are almost always in the $\chi_1$=t state. PEARS uses this information to place these residues before using dead-end elimination and graph decomposition to predict the remaining side chains.
Results & Conclusions
We tested PEARS on a blind test set of 144 antibody models built by ABodyBuilder. Compared to three other side chain predictors, SCWRL, RASP and SIDEpro [4-6], PEARS was the most accurate method, with an average $\chi_1$ accuracy of 79.2% and an average $\chi_1+2$ accuracy of 65.6%. PEARS requires less than ten seconds to model all the side chains of a model structure. Furthermore, PEARS is designed in a similar manner to SCWRL, RASP and SIDEpro, so users can specify individual positions to mutate in an antibody structure.

The advantage of using ABodyBuilder and PEARS is two-fold. Not only are models generated rapidly and accurately, ABodyBuilder provides annotations (e.g. confidence estimate of model accuracy) that can help experimental scientists develop their antibody [1]. With our improved side chain accuracy, these tools will help the antibody design community and improve the results of other tools, such as antibody-antigen docking.

![Figure 1. Model structures predicted by PEARS. Certain side chains, such as L116 Tyr, have a unimodal $\chi_1$ angle distribution; more than 80% of the time, its $\chi_1$ angle is in the g-state. PEARS is suited toward predicting these types of side chains, especially on model structures (grey backbone).](image)

References
1. Leem, J. et al. mAbs, 8(7), 1259-1268 (2016).
79 - Predicting the total hydrophobic surface area of protein structures from sequence

Robbin Bouwmeester¹, Erik van Dijk¹, Jaap Heringa¹, Sanne Abeln¹

¹Computer Science Department, Centre for Integrative Bioinformatics (IBIVU), VU University, Amsterdam, Netherlands

Abstract

We show that predicting the total solvent accessible hydrophobic surface area for a protein structure, given only the amino acid sequence, remains a challenging problem. We develop a simple method based on three global sequence features that outperforms existing methods. Our results suggest that basic physical constraints determine the total amount of hydrophobic surface area for a folded protein.

Introduction

Proteins tend to bury hydrophobic residues inside their core during the folding process. Nevertheless, for many proteins a small fraction of the hydrophobic residues remains on the surface. These hydrophobic patches often play an important functional role, for example in protein-protein interactions, ligand binding and interactions with the membrane. Moreover, knowing the total hydrophobic surface area provides information on aggregation propensity and the propensity to interact with large hydrophobic surfaces in experimental setups.

There are several methods that, from the protein sequence, predict whether a particular residue will be exposed to the surface. Hydrophobic residues will typically be predicted as buried by such methods, following the general trend. Hence, predicting the total solvent accessible hydrophobic surface from the amino acid sequence remains a challenging problem (Fig 1 A-C).
Figure 1. Predicting the exposed surface area for a specific residue is a very different task from predicting the total hydrophobic surface area. The total hydrophobic surface area (A, B: yellow for strongly exposed residues) is vastly underestimated (A, C: blue for underestimated residues) if we simply add the contribution for the hydrophobic residues from single residue based prediction methods (in this case NetSurfP). D: This benchmark shows the ratio of correctly predicted HSA as a function of the error thresholds are shown for several methods; the evaluation was performed on 97 proteins from the test set, that were deposited after 2011 to the PDB, clearly showing the three feature method (TFM in magenta) outperforms the approach of summing hydrophobic contributions from single residues surface prediction methods (e.g. SANN, NetSurfP, SPINEX).

Methods

In order to predict the total surface area from the protein sequence we take two approaches: 1) we took existing single residues surface area prediction methods, including NetSurfP [1], SPINEX [2] and SANN [3], and summed the total surface area for the hydrophobic residues. 2) we built a very simple method that predicts the total amount of hydrophobic surface area based on only three features: the protein sequence length, the number of hydrophobic and hydrophilic amino acids as features. We benchmark the approach against each other, using a set of structure based annotations, and taking care none of the methods have been trained on the final test set.
Results & Conclusions

We show that the simple method based on the three global sequence features predicts the total hydrophobic surface area more accurately than methods for single residue surface area prediction (Fig 1 D). The simplicity of the presented model - which does not take evolutionary conservation into account - indicates that basic physical constraints determine the total amount of hydrophobic surface. Moreover, taking these simple physics based features into account may help to improve the accuracy of other surface based prediction methods; we show for example that our model can be used to improve the accuracy of accessible surface area prediction per residue.

References


81 - Scalable Data Analytics of the PDB with the MacroMolecular Transmission Format (MMTF) and Big Data Technologies

Peter W Rose\(^1\), Anthony R Bradley\(^1\), Antonín Pavelka\(^1\), Alexander S Rose\(^1\), Yana Valasatava\(^1\), Jose M Duarte\(^1\), Andreas Prlić\(^1\)

\(^1\)University of California, San Diego

Abstract

Advances in experimental techniques have led to an explosion in both the number and size of macromolecular structures in the Protein Data Bank (PDB). For this reason, the transfer and parsing of macromolecular data has become increasingly time-consuming. In this work we present the Macromolecular Transmission Format (MMTF), a new compact, extensible macromolecular file format. MMTF offers over 75% compression over mmCIF, and is over an order of magnitude faster to parse than the standard mmCIF format. We describe the new MMTF format, its Application Programming Interface, and demonstrate its use with Big Data Frameworks to enable nearly interactive data analytics on the PDB archive.

Introduction

The MacroMolecular Transmission Format (MMTF) (http://mmtf.rcsb.org/) \([1]\) was designed with three major aims. First, to minimize data size and network transfer times, the format applies custom encoding and compression techniques \([2]\). Both a lossless all atom and a lossy representation with reduced level of detail (C-alpha of polypeptides, phosphate backbone of polynucleotides, and all atoms of ligands) are available. Second, the data files should be fast to parse. Third, MMTF was designed to be self-contained, extensible, and interoperable. As a binary, machine-readable format, the preferred access to MMTF is through its APIs, provided in several programming languages. In this work we show that MMTF enables high-performance structural bioinformatics calculations.

Methods

MMTF is a compact and extensible file format. Compression is enhanced through columnar encoding and custom compression strategies. Dictionary encoding is used to
encode repeating residue, bond and atom information. Delta, run-length, and recursive index encoding strategies are used to compress data arrays, such as atomic coordinates and other atomic properties [2]. Decoder libraries are available for Java, JavaScript, Python, C, and C++. The PDB is available as individual MMTF files and as Hadoop Sequence files for parallel processing. The data files are updated weekly (http://mmtf.rcsb.org/download.html).

Results & Conclusions

In Figure 1 we show that MMTF files are loaded significantly faster than standard mmCIF and PDB file formats.

![Figure 1. Comparison of BioJava [3] load time for the PDB archive (~127,000 entries) using different file formats on a Mac Mini (single core).](image)

In addition, file size (gzipped) for the entire PDB is reduced from 34.1 GB for mmCIF files to 8.1 GB for MMTF-full files and 0.9 GB for MMTF-reduced files.

Using the Apache Spark distributed parallel computing framework, the PDB can be processed in parallel ranging from laptops, workstations, to compute clusters. We show how Big Data operations such as Filter, Map, Reduce, and Broadcasting can be used to implement structural bioinformatics workflows. Further, extracting structure and sequence information into Dataframes enables parallel SQL queries.
Acknowledgements

This project was supported by the NCI of the NIH under award number U01 CA198942.

References


82 - Sensitive and efficient topology-independent structural alignment

Antonín Pavelka¹, Andreas Prlić¹, Peter W. Rose¹

¹University of California San Diego

Abstract

A new topology-independent algorithm for protein structural alignment is presented. The accuracy of the algorithm is compared with the topology-independent CLICK [1] and the sequence-based rigid jFatCat [2] on randomly paired CATH nonredundant S20 domains. The new algorithm always reveals a larger or equal number of alignments with TM-score above a fixed but arbitrary threshold, and 2.5 times more alignments than CLICK at the threshold of 0.5.

Introduction

Structural alignment can be used in studies of homology modeling, protein folding, protein design, and protein evolution [3]. Two proteins can have similar structure even if the corresponding blocks are ordered differently in their sequences, e.g., in case of circular permutation or domain swapping. To take residue connectivity into account and yet allow for arbitrary block shuffling, sequence fragments of a fixed length are assembled into an alignment based only on the similarity of their structures and relative orientations in both structures.

Methods

All sequence fragments that are close in space within a structure are paired into so-called biwords. Alignment is then built starting at each matching pair of biwords separately, by adding matching pairs of words one by one in a greedy fashion. A word pair \((a_1, b_1)\) is proposed for addition if \([a_1, a_2], [b_1, b_2]\) are matching biwords and \([a_2, a_3], [b_2, b_3]\) are matching biwords whose words are already included in the alignment and the structural similarity of \([a_1, a_2], [b_1, b_2]\) is a minimum among all such possible pairs of biwords, where \(a_1, a_2, a_3\) are words from one structure and \(b_1, b_2, b_3\) from the other. The word pair is added only if the residue pairing it defines does map any residue differently from the pairing defined by the alignment.
QCP superposition [4] is used for each large enough alignment to orient the structures. New matching words are then identified utilizing their proximity in the aligned structures. A residue pairing is derived from the matching words and the process is repeated once more starting by superposition.

Biwords are represented as vectors that are stored in multidimensional grid to allow for efficient retrieval of similar vectors, which can be especially useful for database searches. Similarity of a biword pair suggested by vectors is refined by QCP superposition.

TM-score is used to measure alignment quality for all three methods, using the number of residues in the smaller structure for normalization. Whole words are required to match in the aligned structures to avoid isolated residue pairs.

**Results & Conclusions**

TM-scores computed from aligned structures provided by each method are shown in Figure 1. For any TM-score threshold, the new algorithm always finds equal or greater number of similarities than CLICK or jFatCat. For the threshold 0.5, the algorithm detects 2.5 times more similarities than CLICK (176 vs. 71) and 3.8 times more than jFatCat (176 vs. 46). Figure 2 shows an alignment produced by the new algorithm with TM-score 0.59 (vs. 0.25 and 0.35 by CLICK and jFatCat). The computation takes 1.1 seconds per alignment on average using an ordinary notebook.
FIGURE 1. TM-scores for 10,000 random pairs of CATH S20 domains.

FIGURE 2. Alignment of CATH domains 3v2yA01 (blue) and 2x0cA01 (red) generated by the new algorithm. The aligned domains are shown in two orientations.

Acknowledgements

This project was supported by the NCI of the NIH under award number U01 CA198942. Compute resources were provided by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.
References


84 - Dewetting the intracellular water pocket of the human connexin 26 hemichannel via a water-selective repulsive potential

Felipe Villanelo Lizana$^{1,2}$, José Antonio Garate$^{1,2}$, Tomas Perez-Acle$^{1,2}$

$^1$Computational Biology Lab, Fundacion Ciencia & Vida, Santiago, Chile, $^2$Centro Interdisciplinario de Neurociencias de Valparaiso, Universidad de Valparaiso, Valparaiso, Chile

Abstract

Connexins (Cxs) are eukaryotic transmembrane that form hydrophilic channels composed of six Cx monomers, called hemichannels (HCs). The extracellular docking of two opposed HCs, connecting the cytoplasm of two adjacent cells, form a gap junction channel (GJC). HCs and GJCs are expressed in most human tissues being involved in several functions such as cellular differentiation, electrical synapsis and the immune response. These channels allow the passage of solutes up to 1 kDa. Of note, HCs and GJCs exhibit voltage-dependent gating in response to transmembrane and transjunctional potential differences, respectively. Two protein regions flanking the pore are related to gating: the NTH, located to the intracellular side, and the parahelix (PH), facing the extracellular, are related to the fast and the slow gating, respectively$^1$.

The most studied member and the only source of structural information is the human Cx26 (hCx26)$^2$. We recently described an intracellular water pocket, the IC pocket, that is present in each hCx26 monomer$^3$. This pocket is located behind the NTH and it is filled with structural water molecules entering from the pore lumen. Mutations disturbing the IC pocket in both hCx26 and mouse Cx32, hinder channel function, either losing voltage-dependence or changing ionic conductance$^{3,4}$. In accordance, we have hypothesized that IC pocket water volume could modulate the NTH, therefore affecting channel gating. In this work, we asses this hypothesis by implementing a water-selective repulsive potential (WSRP) to dewet the hCx26 IC pocket.

Methods

Molecular dynamics (MD) were performed using the NAMD software package with the CHARMM36 forcefield. An atomistic model of hCx26 HC was prepared adding KCl for
system neutralization and to achieve physiological conditions, as reported elsewhere\textsuperscript{3}. After 100 ns of initial equilibration, the final frame was used as starting point for further MD. To achieve IC pocket dewetting, we created a WSRP mimicking a growing bubble of vacuum in the pocket center. WSRP only affected water molecules and not protein atoms, growing smoothly to a radius of 10 Å. Then, our protocol followed a two-fold approach producing four independent MD simulations of 20 ns for each condition. In the control condition, the WSRP was removed and water molecules could re-enter to the IC pocket. In the test condition, WSRP exerts a repulsive force on both water molecules and the protein atoms. By manipulating the WSRP magnitude, we changed the IC pocket volume to study its effect on the channel function.

Results and conclusion

The application of WSRP dewetted all Cx monomers. When WSRP was removed, all IC pockets become rapidly re-hydrated within the first ns of MD simulation\textsuperscript{3}. According to our hypothesis, by controlling the IC pocket volume we found that the mobility of NTH changes remarkably. In the absence of water molecules, the NTH shows higher mobility compared to the control condition, stabilizing as water molecules enter the pocket. When increasing the volume, the position of the NTH is displaced towards the center of the pore changing the inner radius and affecting ionic conductance. Of note, our structural analyses show that changes in the position of the NTH also modify the position of the PH, providing further evidence for a gating coupling between the NTH and PH\textsuperscript{1,3}.

References

85 - Mutations and Variations in Health and Disease: Protein Interaction Networks and 3D Structure

Franca Fraternali

King’s College London

Abstract

Two parallel efforts are reshaping the microscopic view of the functioning of the human cell and its correlation with healthy and diseased cellular states: a) large-scale proteomics studies with associated functional annotation of the extracted proteins and b) the detection of human variants in healthy individuals and in disease-related cellular states. We aim at merging this information by using in-house developed algorithms for the topological analysis of networks and by extraction of short-loop motifs that highlight associations between proteins and therefore allow for the extraction of disease-related communities of proteins. We extract topologically and functionally connected communities of proteins that are affected by disease mutations and analyse the impact on the underlying 3D-structures to prioritise novel candidates for targeted drug screening.

Introduction

In the last years Systems Biology has provided frameworks to integrate high-throughput biological and clinical data, providing significant insights into some of the fundamental roles of genes and proteins in maintaining a functional cellular state. However, it is still challenging to employ quantitative methods to identify important disease-related relationships between proteins harbouring mutations in their structural domains. We recently developed a method called “short-loop network motif profiling”\(^1,2\) to understand functionality embedded in protein-protein interaction networks (PPIN). The method can be used to design targeted experiments to detect functional phenotypes of proteins in short loops.

Methods

The protein interfaces are identified with POPSCOMP\(^3\) and characterized by analyzing the amino acid composition and conservation, the secondary structure content and the promiscuity of binding. The correlation between interface properties and different per-
Residue flexibility indices are calculated. We use a recently developed platform for 3D structure mapping of SNPs and variants, PinSnps, and we perform a large-scale comparison of impact of mutations on 3D structures from a set of databases featuring common variants, clinically proven pathogenic and benign mutations and cancer-related mutations.

Results and Conclusions

We show that a) short loop network motifs associating proteins with common variants are significantly fewer than loops containing pathogenic variants. b) By generating disease-specific networks like genes involved in different Leukemia sub-types, identify pairs of homologous short loops; for example, kinases (FLT3 and KIT) harbouring pathogenic mutations found in patients with Acute Myeloid Leukemia and localised in identical positions in the 3D-structure. c) We observe clear differences in the distribution of mutation types in different 3D-structure regions, with complementary patterns distinguishing between pathogenic and common variants, suggesting that these properties can be used as input for predictions tools. More generally, we show that 3D PPIN analysis can also help biologists to effectively search for possible targets for disease treatment.

Figure 1. Integrated data for the construction of a comprehensive cross-validated Human PPI Network. Structural Regions considered in the mapping of the different
mutation classes. Enrichment in the different regions of the considered mutations and their significance measured against a random background distribution.

References


86 - Accurate and reliable prediction of relative ligand binding potency in drug discovery - Applications in scaffold hopping transformations

byungchan Kim, David N. LeBard, Yuqing Deng, Lingle Wang, Robert Abel, Dan Wandschneider, Mike Beachy, Richard A. Friesner, Jianxin Duan, Yujie Wu

1Schrödinger GmbH, 2Schrödinger Inc, 3Department of Chemistry, Columbia University

In the past few years successful applications of free energy calculations in drug discovery projects are emerging. A wide range of application areas such as protein-ligand binding, covalent ligand binding, fragment-based optimization, antibody optimization, protein-protein binding and protein stability have been published. However, the existing FEP methods can only be used to calculate the relative binding free energies for R-group modifications or single-atom modifications, and cannot be used to efficiently evaluate scaffold hopping modifications to a lead molecule. Scaffold hopping or core hopping, a very common design strategy in drug discovery projects, is not only critical in the early stages of a discovery campaign where novel active matter must be identified, but also in lead optimization where the resolution of a variety of ADME/Tox problems may require identification of a novel core structure. Here we present a method that enables relative protein-ligand binding free energy calculations to be pursued for scaffold hopping modifications. We apply the method to six pharmaceutically interesting cases where diverse types of scaffold hopping modifications were required to identify the drug molecules ultimately sent into the clinic. For these six diverse cases, the predicted binding affinities were in close agreement with experiment, demonstrating the wide applicability and the significant impact Core Hopping FEP may provide in drug discovery projects.
87 - Binding of cationic porphyrins to hemoglobin and cytochrome C by the method of molecular docking

Aram Gyulkhandanyan

Institute of Biochemistry of NAS of Armenia

Abstract

The method of molecular docking has been used to study the binding sites of porphyrins to proteins (hemoglobin and cytochrome C), the effect on this binding of fatty acids has been determined, and possible binding sites of [porphyrin + fatty acid] complexes on the macromolecule of the protein have been studied.

Introduction

In photodynamic therapy of tumors (PDT) as photosensitizers, cationic porphyrins are now widely used [1]. It is known that serum albumin (CA) is one of the main carriers of porphyrins in the blood [2] and that long-chain fatty acids (stearate and palmitate) compete for sites of binding in the protein and this significantly weaken the transport process [3]. It is also well known that two heme-containing proteins - hemoglobin (Hb) and cytochrome C (Cyt C) - perform the most important functions in nature, but to date, many aspects of the binding of these proteins to porphyrins have not yet been studied. The method of molecular docking allows not only the study of sites of porphyrin binding to the protein, but also additionally determine the effect on this binding of other ligands (fatty acids) and to study possible binding sites for [porphyrin + fatty acid] complexes on the macromolecule of the protein. The present work is devoted to the important aspect of research of new properties of these proteins - to studying binding of cationic porphyrins with human hemoglobin and cytochrome C and the effect on this process of fatty acids by molecular docking method.

Methods

The structure of proteins (www.pdb.org) was used for molecular docking with cationic porphyrins, as well as for complexation with stearic or palmitic acids. The molecular models of porphyrin and FAs were built using the ChemBioOffice Ultra 11.0 program (http://software.informer.com/getfree-chembio3d-ultra-11.0/). The docking procedure and visualization were performed by means of the Graphical User Interface AutodockTools1.4.5. program (ADT) (http://mgltools.scripps.edu/), using its AutoGrid and AutoDock functions [4].

Results & Conclusions
By the molecular docking method is shown,

• that porphyrins, fatty acids and their complexes in the hem binding site of the hemoglobin can not penetrate deep into the macromolecule and remains on the surface;

• complexes [porphyrin + fatty acid] bind to the inner central cavity of the hemoglobin macromolecule much more probability than porphyrins or fatty acids, and thereby [porphyrin + fatty acid] complexes can displace porphyrins from the inner cavity of the hemoglobin macromolecule.

FIGURE 1. Docking of complex [porphyrin + stearic acid] (shown in green and cyan, respectively) in the inner cavity of the hemoglobin macromolecule (a). In blue represents 4 molecules of heme. On the right (b) is the positioning of the complex [porphyrin + stearic acid] (shown in green and cyan) with α1β1 chains of the hemoglobin macromolecule.

At the same time, the methods of absorption spectroscopy and small molecules microarrays have shown that if in the presence of fatty acids the degree of binding to porphyrins for serum albumin decreases moderately, then for hemoglobin the degree of binding decreases more significantly. In this regard, the main role of the protein-transporter of cationic porphyrins among blood proteins should be given to serum albumin.

References


88 - Protein S-palmitoylation site prediction using position-specific scores

Tatsuki Kikegawa¹, Tsubasa Ogawa¹, Kenji Etchuya¹, Yuri Mukai¹

¹Department of Electronics, Graduate School of Science and Technology, Meiji University

Abstract

S-palmitoylation is an important post-translational protein lipid modification for controlling protein function and subcellular localization. A new method for predicting protein S-palmitoylation sites using a combination of the position-specific score and a back propagation artificial neural network was developed in this study.

Introduction

S-palmitoylation plays a role in regulating protein function and subcellular localization in response to an external stimulus. S-palmitoylation is catalyzed by the S-palmitoylation enzyme S-palmitoyltransferase protein acyltransferase (PAT). PAT proteins can be classified into three groups depending on the characteristics of their target proteins: those interacting with (i) soluble proteins, (ii) transmembrane proteins, and (iii) proteins modified by other lipids (FIGURE 1) [1]. PAT proteins in each group are thought to recognize different S-palmitoylation motifs. Therefore, classifying the target proteins into three groups is an essential step toward detecting S-palmitoylation motifs.
In the present study, a new method for protein S-palmitoylation site prediction with high accuracy was developed. In the proposed method, S-palmitoylated protein sequences were classified into three groups (soluble proteins, transmembrane proteins, and proteins modified by other lipids), and the characteristics of each group were extracted and applied to a back propagation artificial neural network.

**Methods**

S-palmitoylated protein sequence data were obtained from the UniProt Knowledgebase/Swiss-Prot protein sequence database (URL: http://www.uniprot.org/) and Swisspalm (URL: http://swisspalm.epfl.ch/) by conducting a search with the keywords “Eukaryota” and “S-palmitoyl cysteine”. The data was applied to classify the S-palmitoylated protein data into one of three groups, soluble protein group, transmembrane protein group, and the group of proteins modified by other lipids. Positive and negative sequence datasets of each group included 51 amino acids from the -25th to +25th position relative to the position of the S-palmitoylated and non-palmitoylated Cys residue, respectively. PSSs of each group were calculated.

A method for the protein S-palmitoylation site prediction was developed by a combination of the PSS and BP-ANN. BP-ANN with a three-layered structure was used.
for the prediction of S-palmitoylation sites. PSSs of each group for n residues from S-palmitoylation site (2 ≤ n ≤ 25) were applied to the input nodes. In each group, 5-fold crossvalidation was performed while the calculation area was changed for the purpose of excluding the range with a possibility of over fitting. PSSs calculated according to position-specific amino acid propensities were assigned to the input layer referring to the query sequence. A score in the output layer greater than 0.5 indicates that the Cys residue in the query sequence is predicted as an S-palmitoylation site.

Results & Conclusions

The S-palmitoylation prediction accuracy was improved by applying this three-group classification of the dataset. The accuracy in this study was better than the prediction methods reported previously. In the soluble proteins and proteins with other lipids groups, S-palmitoylation prediction was performed with high accuracy by the combination of the PSS and BP-ANN. In the transmembrane protein group, the prediction accuracy was less than 80%, but it was improved by considering the protein topology. The proposed method for S-palmitoylation prediction can be a powerful tool to comprehensively find candidates not only for unknown S-palmitoylated proteins but also for unknown S-palmitoylated sites on known palmitoyled proteins.

References

Molecular docking studies can shed light into the molecular determinants of ligand binding. When no experimental structures are available, docking can also be applied to homology models. Given that the model quality, especially in the binding site, greatly affects the accuracy of docking predictions, development of strategies able to include protein flexibility may help in the effective use of docking to homology models. In this work, ligand binding of structurally different ligands to the mouse Aryl hydrocarbon Receptor (mAhR) homology model is analyzed.

Introduction

AhR is a ligand-dependent transcription factor that responds to exogenous and endogenous chemicals with the induction of gene expression and production of diverse biological and toxic effects. The mechanism is initiated by ligand binding to the AhR, which is present in the cytosol as a multiprotein complex, and the PAS-B domain acts as ligand binding domain (LBD).

The most “classical” AhR ligands are the exogenous halogenated aromatic hydrocarbons (HAHs) (including the most toxic 2,3,7,8-tetrachlorodibenzo-p-dioxin, TCDD), polycyclic aromatic hydrocarbons (PAHs), and PAH-like chemicals; but also “non-classical” natural, endogenous and synthetic AhR ligands with diverse structure and physico-chemical characteristics have been identified.

Given that AhR shows different biological responses for different ligands, understanding the molecular determinants of binding could help in elucidating the mechanism of action associated to each ligand.

Methods
No experimental information is available for the AhR LBD but since last years the X-ray structures of many homologous PAS domains have been determined, in particular the PASB of the Hypoxia-Inducible Factor 2α (HIF2α) shows 30% of sequence identity with the mAhR LBD.

10 homology models of the mAhR LBD were developed using MODELLER, each one based on a different HIF2α template structure in order to describe the flexibility of the binding cavity. All the ligands and solvent molecules in the HIF2α internal cavity were maintained during the modeling and refinement stages.

An ensemble docking approach (using Glide XP) based on ligand docking to the 10 LBD models was used to include receptor flexibility.

Selected binding poses were subjected to Molecular Dynamics simulation (10–20ns). The binding free energy ($\Delta G_{\text{bind}}$) was calculated by the MM-GBSA approach implemented in AMBER, the most stabilizing contributions were obtained by per-residue decomposition analysis.

**Results & Conclusions**

The set of 12 ligands here analyzed is composed by classical and non-classical AhR ligands with diverse structures and physico-chemical characteristics.

Protein flexibility and plasticity were introduced in both the ensemble docking to multiple homology models and the post-docking MD simulations to obtain an accurate description of binding of such diverse chemicals within the binding cavity. The $\Delta G_{\text{bind}}$ analysis, performed to identify the most relevant residues for each pose stabilization, allowed to discriminate the molecular determinants for binding of the different ligands.

Three main arrangements within the binding cavity were identified (Figure 1): hydrophobic molecules interact mainly with residues located at the bottom of the cavity (TCDD); others occupy the whole cavity and in some cases are stabilized by hydrogen bonds with residues in the middle of the cavity (FICZ); the smallest and polar ligands stay at the entrance of the cavity (leflunomide).
These poses will be validated by a mutagenesis study focused on the key stabilizing residues predicted by the computational protocol here presented.

**FIGURE 1.** The mAhR models, the docking poses of three representative ligands, and the residues that mainly contribute to ΔGbind are shown.

**References**

90 - An exploration of the structural interactome of Rac1

Marijne Schijns1,2, Jessica Andreani1, Raphael Guerois1

1Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette cedex, France, 2Utrecht University, Utrecht, the Netherlands

Abstract

Rac1 is a member of the small GTPase superfamily. Members of this family are characterized by their interaction with effector protein partners when in the GTP-bound state. Rac1 has been implicated in regulation of cell motility and the actin network through a large number of effector proteins. To better understand the role of Rac1, and the effect of mutations in Rac1, models of Rac1 in contact with its partners are essential. Rac1 exposes two adjacent apolar patches in its GTP-bound state, termed its “switches”. Burial of these apolar patches is highly favored by docking methods and typically biases docking outcomes. Though previous research has shown that the switches are often involved in Rac1 interactions, this is not always the case. This raises two main questions when docking Rac1 with one of its partners: (i) how to correctly identify models that place the partner away from the switches on Rac1 and (ii) assuming that binding to these switches is correct, how to remain critical of the docking solutions?

To address these questions, we consider a set of binding partners of Rac1 for which structural and evolutionary data of both Rac1 and the partners are available. These existing structures are retrieved to serve as benchmark for Rac1 binding motifs, which is used to better understand and evaluate docking results for Rac1 complexes.

This results in a set of models of Rac1 and its partners, elucidating the modes of interaction of Rac1 and providing a basis of its structural interactome. Additionally, it may give more insight in the methodology surrounding docking and scoring concerning proteins with apolar surface areas.
91 - Do trends in biomacromolecular structure quality inspire optimism?

Vladimír Horský¹, Veronika Bendová¹,², Radka Svobodová Vařeková¹, Sameer Velankar³, Jaroslav Koča¹

¹National Centre for Biomolecular Research, Faculty of Science, and CEITEC - Central European Institute of Technology, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic, ²Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Kotlářská 267/2, 611 37 Brno, Czech Republic, ³Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom.

The general availability of data about biomacromolecular structures is a key success of modern life sciences. In fact, 13 Nobel prizes were awarded for research based on this data [1]. Acquired structures are stored in databases, the most prominent one in structural biology being the Protein Data Bank (PDB) [2]. Not all that glitters is gold, however. Structure errors have caused retractions of articles from reputable journals [3]. In reaction, scientific community began developing various tools suited for validating of biomacromolecular complexes [4-9]. PDB began offering validation reports that enabled its users to assess wide range of quality criteria of each individual structure [10-11]. Therefore, we became curious if their emergence had positive influence of quality of newly submitted structures. And since ligands do not enjoy as much care as biomacromolecules, we were also curious about trends in ligand quality.

Wide scale analysis of trends in quality and size of biomacromolecules and their ligands has therefore been carried out by our team. 88 factors have been considered. Structure metadata and quality data came from PDB, while ligand quality data have been sourced from our own database ValidatorDB [9].

Some trends were expected to exist (e.g., newer structures have better quality), while the existence of others was a surprise (e.g., ligand quality is stagnant at best, currently utilized structure validation methods do not validate ligands well). Explored trends are available in the ValTrendsDB database (ncbr.muni.cz/ValTrendsDB).

References

1. Protein Data Bank Europe web pages: Structural biology related Nobel Prizes: http://www.ebi.ac.uk/pdbe/docs/nobel/nobels.html


92 - Dexterity: A framework to use a smartphone as a 3D wand

Jenny Vuong¹, Benedetta Frida Baldi², Christopher J. Hammang², Seán I. O'Donoghue¹,²

¹Data61, CSIRO, Sydney, NSW, Australia, ²Garvan Institute of Medical Research, Sydney, NSW, Australia

Advances in web-based 3D graphics are making 3D datasets more accessible to a rapidly growing community of both specialists and lay people. A key challenge common to almost all 3D applications is onboarding, since the standard methods used for 3D controls (e.g., rotate, translate, zoom, etc.) can be difficult for new users to learn. Additionally, these 3D controls often change with different input devices, requiring the user to learn multiple control combinations that trigger the same functionality. Another key challenge is depth cueing. Both challenges can be partly addressed using dedicated control devices, such as a 3D wand. However for web applications, it is generally best to avoid depending on the availability of such devices, as these are difficult to acquire - quite often only purchasable at specialist shops - and/or are associated with high costs.

Thus we propose Dexterity, a web-accessible prototype framework that allows a smartphone to be used as a 3D wand to control 3D objects. Using standard JavaScript libraries and web-based functionalities (e.g. Socket.io for WebSockets, deviceorientation JavaScript event, etc.), we implemented a simple web-based application that can be accessed using a mobile browser on any smartphone. Given the wide availability of smartphones, Dexterity thus provides a generic and cost-effective solution that helps address the challenges of onboarding and depth perception.
93 - Structural and functional analysis of alternative microexons of proteins observed in RNA-seq studies.

Matsuyuki Shirota$^{1,2,3}$

$^1$Graduate School of Medicine, Tohoku University, $^2$Tohoku Medical Megabank Organization, Tohoku University, $^3$Graduate School of Information Sciences, Tohoku University

Abstract

Alternative splicing involving very short (3~15 nt) exons, referred to as microexons (MEs), have been shown to be important for neuronal development and pathogenesis of psychiatric diseases. In this study, the cell-type and genotype specific regulations of such MEs were identified and their impact on protein function was inferred by using protein 3D structure.

Introduction

The advancement of high-throughput RNA sequencing (RNA-seq) has enabled us to discover various junctions between exons within a mature mRNA, some of which may be observed in specific types of tissues, cells or developmental stages, or in the individuals having particular genotypes, due to alternative splicing (AS). AS events involving very small exons ranging from three to fifteen nucleotides, or MEs, will introduce deletions or insertions of one or a few amino acids to the protein sequence, which can cause changes in protein structure and function. Although much efforts has been made for inferring the impacts of missense SNVs based on available protein 3D structures, understanding and predicting the consequences of the AS of such MEs from a perspective of protein structure has not been well performed.

Methods

We used two publicly available RNA-seq data sets; one from single cell RNA-seq of human brain and the other from lymphoblastoid cells of 462 individuals from 1000 genomes project. The level of differential joining and skipping of alternative MEs in various proteins was quantified by percent-spliced in (PSI) by using vast-tools. Briefly, the fastq files for each project were mapped by Bowtie program to a database of sequences consisting of exon-exon junctions either including or excluding an ME. PSI
was calculated as the ratio of the number of reads including the ME and the sum of the reads both including and excluding the ME. The expression levels of genes were also quantified by using TopHat program.

**Results & Conclusions**

Each of the single cell RNA-seq data of human brain was classified into cell subtypes, i.e. neurons, oligodendrocytes, astrocytes, etc., based on the expression levels of lineage specific genes. Those MEs that were previously reported to be up-regulated (frequently included) in neuronal tissues were observed specifically included in the neuron cells, thereby reproducing the previous report in single cell resolution. Inclusion of MEs in mRNAs was observed as a switch-like event among single cells; the PSI of an ME in each cell was either close to 0% (excluded) or 100% (included), which was quite different from the distribution of PSI obtained from RNA-seq of bulk of cells.

For the RNA-seq data of lymphoblastoid cells, genome-wide associations between PSI of MEs and genotypes were tested among 462 individuals. Significant associations were observed for the MEs of three genes, AP1B1, LYRM1 and POLR2J2, with SNVs in the genomic regions neighboring the MEs, suggesting their role as cis-working splicing quantitative trait loci (Figure 1).

In protein 3D structures, most these MEs occur in the surface of the protein, and some of them occur within α-helix or β-sheet, suggesting the structural and functional

**FIGURE 1.** A genome wide association result for an ME at chr22:29725701-29725709 (9) in AP1B1 gene. A) Manhattan plot and B) PSI for each genotype. 

perturbation. Taken together these results indicate that MEs are precisely regulated by genetic factors, and protein 3D structure can be used to shed light on their biological impact.

References

Introduction

When proteins exert their functions, physical contact with other molecules occurs. A close connection therefore exists between their functions and structures. Analyzing the interactions between proteins and small molecules is important to predict the ligands which bind to parts of putative ligand binding pockets. Such analysis demands a fast and efficient method for comparing ligand binding pockets. Methods have been developed for representing a ligand binding pocket with one reduced vector for binding pocket comparison [3]. However, some possible caveats must be associated with existing methods.

1) These methods have no abundant ability of expression because they cannot increase the number of patterns of labels for counting the occurrence frequency of triangles because the occurrence frequency vector becomes sparse and because it is difficult to calculate the similarity properly.

2) The error by which the similarity between similar triangles is regarded as completely dissimilar occurs when a ligand binding pocket is converted into triangles because the similarity between triangles is not considered beyond the bin.

To overcome these difficulties, we developed a novel method [4, 5].

Methods & Datasets

Each binding pocket is described by an ensemble of all triangles consisting of three Cα atoms of amino acids in the pocket, in contrast to our previous study (Existing Method) which considered limited types of amino acids at a time (Fig. 1). Each triangle vertex is labeled with one amino acid of 20 types. Regarding the triangle edges, we classified edges into six classes and assigned them Roman numerals (I, II, III, IV, V, and VI). We regarded triangles as identical if they have the same label, with regard to their vertices and edges, considering all six ways of superposition. Given these conditions, we listed
all possible triangle types, i.e. 295,240 types in total. Then, we defined the similarity among all triangles which is producible under our labeling method, and calculated vector representation of triangle type ($X$) using multidimensional scaling (MDS) with randomized eigenvalue decomposition. Using $X$, we defined a reduced representation of pocket as $w(=Xn)$.

**FIGURE 1. Schematic diagram of a vector representation of a ligand-binding pocket.**

We examined two datasets for this study: i) *Ito138*. This difficult dataset comprises pocket pairs that share the same type of small molecules made from 138 known ligand binding pockets. ii) *APocS3*. This relatively easy dataset comprises 37,956/26,527 pairs for Subject (binding similar ligand)/Control (not similar) dataset.

**Results & Conclusions**

First, we investigated the effects of change of the scheme from Existing Method to New Method (in Fig. 1) for defining the similarity between pockets using *Ito138* dataset. From the view point of ROC curves, the new method outperforms both FuzCav-like and our previous method, PoSSuM-like (Existing Method) [2], in terms of AUC. We confirmed that these improvements in performance derived from the new scheme in which we took account of the similarity between triangle types.

Second, we compared our new method with an existing fast sequence order-independent structural alignment method: APoc [1] using *APocS3* dataset. We found that the new similarity definition also outperforms APoc.
Moreover, the computational times for calculating similarity of randomly selected pocket pairs suggested that this new method can provide similarities faster than APoc. In addition, we showed, for some examples, that we can handle pocket conformational variation associated with the ligand conformation change contrary to APoc.

References


95 - Inference of functional states from conformational changes in protein complexes

Markus Gruber¹, Manfred Sippl¹

¹University of Salzburg

Abstract

The association of chains into complexes is a common way for proteins to augment their capabilities. In particular, several examples are known, where a difference in the association of the chains of the same protein complex reflects different functional states of the molecule. Here, we compare the atomic models of identical and homologous pairs of protein complexes to automatically identify differences in the association of their constituent chains. Results suggest that changes in the association of the chains are strongly tied to functional states of protein molecules and may prove useful to infer functional insights.

Introduction

The association of single protein chains into complexes augments the diversity of proteins and opens new ways how proteins may interact and function. As the number of known protein complexes increased steadily in recent years, more and more of their enhanced functionalities obtained through multimerization became apparent. In particular, differences between the association of the chains in the same complex appear to be linked to distinct functional states of a molecule (Figure 1). By comparing atomic models of identical and related protein complexes, differences in the association of the chains can be recognized automatically. Advanced structure matching techniques [1,2], which are able to align atomic models of protein complexes, allow us to perform a large scale analysis to identify differences in chain associations in order to understand if these differences are indeed linked to functional states of molecules.

Methods

For the comparison of protein complexes and their constituent chains we use the structure alignment program TopMatch [1]. TopMatch is able to align protein complexes as a whole thereby capturing the relative spatial arrangement of the constituent chains [1,2]. From the alignment of two complexes we can extract the chain mapping - which
chain in complex one belongs to the corresponding chain in complex two - to align each of the obtained chain pairs independently. By comparing the similarity of the complex alignment and the sum of the similarities of each chain alignment we can measure the amount of change in the chain associations.

Results & Conclusions

We measured and quantified the differences in the chain associations of identical and related proteins in two large datasets. We compared (1) complexes originating from the same crystal of an X-ray structure determination experiment and (2) homologous complexes with at least 50\% sequence identity independent of their experimental origin. Complexes were obtained by extracting all biological assemblies with more than one chain from all protein structures available from the PDB [3]. Our results include known examples of proteins having distinct functional states as well as cases where little is known about what has caused the difference the association of the chains. The latter are prime examples for further experimental investigations about their functioning.
96 - Structural characterization of the IC pocket in the human Cx50 hemichannel

Claudia Pareja-Barrueto\textsuperscript{1,2}, Jose Gomez\textsuperscript{1,2}, Felipe Villanelo\textsuperscript{1,2}, Peter Minogue\textsuperscript{3}, Viviana Berthoud\textsuperscript{3}, Eric Beyer\textsuperscript{3}, Tomas Perez-Acle\textsuperscript{1,2}

\textsuperscript{1}Computational Biology Laboratory (DLab), Fundacion Ciencia y Vida, Santiago, Chile, \textsuperscript{2}Centro Interdisciplinario de Neurociencia de Valparaiso, Valparaiso, Chile, \textsuperscript{3}Department of Pediatrics, University of Chicago, Chicago, IL USA.

Previous \textit{in silico} studies, developed in our laboratory, suggest that gating of human Cx26 hemichannels can be modulated by an intracellular water pocket (IC pocket). The presence of a water pocket has also been reported in Cx32 hemichannels. Both Cx26 and Cx32 belong to the \(\beta\) subfamily of connexins. To elucidate whether members of other connexin subfamilies also contain a water pocket, we chose human Cx50 (hCx50), a member of the \(\alpha\) subfamily of connexins. Cx50 is expressed in the eye lens, and several mutations in its gene have been associated with congenital cataracts in humans.

To test for the presence of the IC pocket in hCx50 and its possible functional role, we performed all-atom molecular dynamics simulations on wild type hCx50 hemichannels and on hemichannels formed by hCx50 containing mutations of some of the amino acid residues lining the IC pocket.

Our analyses are focused on the structure and dynamics of water molecules inside the IC pocket and ionic currents across the channel pore. We calculated occupancy, survival probability, and dipole vector orientation of water molecules, and made structural comparisons. In addition, we performed functional mode and essential dynamics analyses to identify relevant collective atomic motions of the hemichannel.

Our results suggest that hCx50 hemichannels contain an IC pocket. Moreover, water occupancy and dynamics imply the presence of highly structured waters that make electrostatic interactions with particular amino acid residues lining the IC pocket. Thus, the existence of an IC pocket is a general feature of hemichannels formed by different connexins. Ongoing studies will evaluate its importance for hCx50 function.

\textbf{Acknowledgements}

This work was supported by Fondecyt grant #1160574, PFB16 Fundación Ciencia para la Vida and ICM-Economía P09-022-F, CINV, Conicyt graduate fellowships #21161628 and NIH grant EY08368. We acknowledge the University of Chicago RCC for access to supercomputing time. Powered@NLHPC: This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02).
97 - Using normal modes analysis to characterize the flexibility of protein tunnels and channels

Pierre Bedoucha¹, Edvin Fuglebakk³, Jan Byska⁴, Nathalie Reuter²

¹University of Bergen, ²University of Bergen, ³Institute of Marine Research in Norway, ⁴Masaryk University

Abstract

Protein 3D structures are tightly related to protein functions. There are withal missing links between protein structure and function and studies have shown that protein dynamics is one of them.

The transport of compounds of various sizes through cell membranes is partly ensured by transmembrane proteins, such as channels or carriers. Their 3D structure is often characterized by an opening providing a route through which ions or small molecules will be traveling. During the transport process these tunnels may change geometry, adjusting their access and permeability to regulate the protein’s function. These structural changes are happening on time scales that are often too long to be captured by molecular dynamics simulations.

We propose to use Normal Mode Analysis (NMA) to investigate the repercussion of protein intrinsic dynamics on the properties of tunnels in proteins 3D structure. NMA has indeed been repeatedly shown to be an efficient and reliable method to describe slow and large amplitude movements in proteins (Fuglebakk et al., BBA, 2015). We combine NMA with the use of CAVER (Chovancova et al., PCBI, 2012), a tool for the analysis and visualization of protein tunnels and cavities. We have developed a computational framework linking CAVER and NMA results. We have validated it on a few proteins for which we have compared the use of coarse-grained (CG) and all atoms representations to model channel flexibility. The use of CG representations implies residue side-chain reconstruction as cavities are highly sensitive to side chain positions so we have also introduced this in our pipeline.

We confirmed that NMA is a fitting mean to investigate tunnel and channel plasticity and thus can be used as an input for investigation tools such as CAVER.

Keywords: protein normal mode analysis, transmembrane proteins, protein tunnels
References


98 - Sensitive and efficient topology-independent structural alignment

Antonín Pavelka¹, Andreas Prlić¹, Peter W. Rose¹

¹University of California, San Diego

Sensitive and efficient topology-independent structural alignment

Abstract

A new topology-independent algorithm for protein structural alignment is presented. The accuracy of the algorithm is compared with the topology-independent CLICK [1] and the sequence-based rigid jFatCat [2] on randomly paired CATH nonredundant S20 domains. The new algorithm always reveals a larger or equal number of alignments with TM-score above a fixed but arbitrary threshold, and 2.5 times more alignments than CLICK at the threshold of 0.5.

Introduction

Structural alignment can be used in studies of homology modeling, protein folding, protein design, and protein evolution [3]. Two proteins can have similar structure even if the corresponding blocks are ordered differently in their sequences, e.g., in case of circular permutation or domain swapping. To take residue connectivity into account and yet allow for arbitrary block shuffling, sequence fragments of a fixed length are assembled into an alignment based only on the similarity of their structures and relative orientations in both structures.

Methods

All sequence fragments that are close in space within a structure are paired into so called biwords. Alignment is then built starting at each matching pair of biwords separately, by adding matching pairs of words one by one in a greedy fashion. A word pair \((a_1, b_1)\) is proposed for addition if \([a_1, a_2], [b_1, b_2]\) are matching biwords and \([a_2, a_3], [b_2, b_3]\) are matching biwords whose words are already included in the alignment and the structural similarity of \([a_1, a_2], [b_1, b_2]\) is minimum among all such possible pairs of biwords, where \(a_1, a_2, a_3\) are words from one structure and \(b_1, b_2, b_3\) from the other. The
word pair is added only if the residue pairing it defines does not map any residue differently from the pairing defined by the alignment.

QCP superposition [4] is used for each large enough alignment to orient the structures. New matching words are then identified utilizing their proximity in the aligned structures. A residue pairing is derived from the matching words and the process is repeated once more starting by superposition.

Biwords are represented as vectors that are stored in a multidimensional grid to allow for efficient retrieval of similar vectors, which can be especially useful for database searches. Similarity of a biword pair suggested by vectors is refined by QCP superposition.

TM-score is used to measure alignment quality for all three methods, using the number of residues in the smaller structure for normalization. Whole words are required to match in the aligned structures to avoid isolated residue pairs.

**Results & Conclusions**

TM-scores computed from the aligned structures provided by each method are shown in Figure 1. For any TM-score threshold, the new algorithm always finds equal or greater number of similarities than CLICK or jFatCat. For the threshold 0.5, the algorithm detects 2.5 times more similarities than CLICK (176 vs. 71) and 3.8 times more than jFatCat (176 vs. 46). Figure 2 shows an alignment produced by the new algorithm with TM-score 0.59 (vs. 0.25 and 0.35 by CLICK and jFatCat). The computation takes 1.1 seconds per alignment on average using an ordinary notebook.

**FIGURE 1.** TM-scores for 10,000 random pairs of CATH S20 domains.
FIGURE 2. Alignment of CATH domains 3v2yA01 (blue) and 2x0cA01 (red) generated by the new algorithm. The aligned domains are shown in two orientations.

Acknowledgements

This project was supported by the NCI of the NIH under award number U01 CA198942. Compute resources were provided by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

References


99 - A Web Based 3D Visualization Solution for Large Scale Data

Chao-Chun Chuang¹, Nan-yow Chen¹

¹National center for high-performance computing, Taiwan

Abstract

3D visualization is helpful for understanding complex neuron structure. It also can help scientists share their visualization neuroimaging data and make their representation easier to understand. In most cases, scientists downloaded these big neuroimaging data and then visualized on a local desktop environments with 3D visualization tools. Thus, we demo a web-based remote visualization architecture for these big neuroimaging data (10 GB), and take Fly 3D viewer as an example. Fly 3D viewer uses advanced 3-layer web technologies (Figure 1), web server¹ (WebSockets, and Canvas), visualization server² (VNC and VirtualGL) and user client³ (HTML5) for interactive online 3D surface, skeleton and volumetric neuroimaging data in any web browser environments without requiring any browser plugin (Figure 2). The first layer, web server communicates with the second layer visualization server by WebSockets and Canvas. Then third layer user client can view 3D neuroimaging data from web page constructed by HTML5. Thus, with powerful visualization server, users can easily interactive the 3D visualization neuron imaging by web browser in a lightweight computer resource. Finally, this architecture can be applied to other visualization software⁴⁻⁶ (Vaa3d, Peng et al., 2010; NNG, Lin et al., 2012) from different research groups as well as manage bio-images with deeper neurological insight.

![Diagram](image)

**FIGURE 1.** Three layer architecture web technologies, web server (WebSockets, and Canvas), visualization server (VNC and VirtualGL) and user client (HTML5) for display large scale data.
FIGURE 2. Fly 3D viewer for visualization and interactive online 3D surface, skeleton and volumetric neuroimaging data in any web browser environments

References
100 - LIBRA-WA: a web application for ligand binding site detection and protein function recognition

Daniele Toti¹, Le Viet Hung², Valentina Brandi¹, Valentina Tortosa¹, Fabio Polticelli¹,³

¹Department of Sciences, Roma Tre University, 00146 Rome, Italy, ²Department of Sciences and Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam, ³National Institute of Nuclear Physics, Roma Tre Section, 00146 Rome, Italy

Abstract

Recently we described LIBRA, a stand-alone software tool that detects active and ligand binding sites in an input protein structure. LIBRA’s performance, assessed using the LigaSite database as a test set, was successful in the vast majority of the cases but its scoring scheme and output presentation evidenced room for improvement. To solve these issues the web application LIBRA-WA, based on an improved LIBRA engine, has been developed and made accessible online. LIBRA-WA features a novel scoring scheme, which consistently improves the performance of LIBRA, an improved output presentation and a measure of confidence to help identify false positive hits. LIBRA-WA is available at: http://www.computationalbiology.it/software.html

Introduction

We recently developed and described LIBRA, a graph theory-based software tool that, given a protein structure model, detects the presence of active/ligand binding sites¹. LIBRA, based on the comparison between an input protein structure and a database of more than 170,000 known ligand binding sites, detects subsets of residues in the input protein structure which display stereochemical similarity with known binding sites. Results are returned in the form of structural alignments with known binding sites and relative ligands and ranked based on RMSD or alignment size. Extensive tests carried out on the LigaSite set² indicated that LIBRA was able to identify the correct binding/active site in 90% of the cases. However, the identified correct site ranked first only in 80% of the cases, making it difficult to identify false positive hits. To improve this performance, a novel version of LIBRA with a modified scoring system and a number of additional features has been developed and implemented and tested.

Methods
The main improvement in LIBRA-WA is its novel scoring system, which takes advantage of a SMILES-based clusterization of the more than 20,000 ligands stored in the application’s database. LIBRA-WA, for each alignment record, provides a score combining the alignment’s clique size, RMSD value, and number of ligand hits falling into the same cluster. Besides, the novel detection algorithm allows the identification of binding sites hosted at the interface between different subunits by considering, for each residue of a given chain, its neighbors from different chains within a distance threshold. This has been done by indexing the input protein residues via n-dimensional indexes called KD-Trees. LIBRA-WA results can be also graphically displayed in three-dimensions via the Jmol HTML5 plug-in.

Results and Conclusion

The effectiveness of LIBRA-WA has been evaluated using the LigaSite set of 374 apoproteins as a test set. LIBRA-WA finds the biologically relevant ligand/binding site in ~96% of the cases. Most importantly, the correct ligand/binding site ranks first in almost all of the cases (94%).

This performance is in many cases independent from the sequence/structure similarity between the input protein and the holo protein hosting the known binding site. An example is the E. coli adenylate kinase (apoprotein PDB code 4AKE) which, upon ADP binding undergoes a large conformational change. As a consequence, LIBRA-WA does not identify the ADP binding site of the corresponding holo protein (PDB code 2ECK). However it correctly identifies the ADP binding site in the input protein because of the local structural similarity with the ADP binding site of the human kinesin-8 motor domain (PDB code 3LRE) which shares a mere 11% sequence identity with the adenylate kinase and no significant similarity in the overall protein topology (Figure 1).
FIGURE 1.

References

