The 8th Structural Bioinformatics and Computational Biophysics Meeting
An ISMB satellite meeting, Long Beach California

July 13-14, 2012

Scientific Committee:

Philip E. Bourne - Head (University of California), Rafael Najmanovich (Université de Sherbrooke), Ilan Samish (Weizmann Institute of Science), Ada Yonath (WIS, Nobel Laureate Chemistry 2009), Adam Godzik (SBMRI), Arthur Olson (Scripps), David Goodsell (Scripps), François Major (Université de Montreal), Olivier Lichtarge (Baylor), Chris Sander (Memorial Sloan-Kettering Cancer Center), Russ Altman (Stanford University)

Organizing Committee:

Philip E. Bourne - Head (University of California), Rafael Najmanovich (Université de Sherbrooke), Ilan Samish (Weizmann Institute of Science), Matthieu Chartier (Université de Sherbrooke)
The Warren DeLano Structural Bioinformatics and Computational Biophysics Award

The DeLano Family is pleased to be working with 3DSIG to present The Warren DeLano Structural Bioinformatics & Computational Biophysics Award.

Warren followed his three passions, science, computing and open source to bring PyMOL into the world. Warren devoted his career to the refinement of PyMOL’s code and the technical support of its use. His contributions have furthered understanding and facilitated scientific progress in biophysics, biochemistry and molecular biology. PyMOL’s open-source accessibility has allowed students to visualize implications of theory, has provided broke post-docs with the ability to publish images regardless of budgetary constraints and has afforded the lone computational biologist the means to build a specific tool to better understand a structure or process.

The DeLano Family encourages you to follow your passions. Further, we hope some of you are inspired by the product of Warren’s work and his devotion to an “open source” methodology. We feel that Warren lives on in the code he created - but so-too in the manner in which he shared his brainchild; his life’s work stands as testament to the value of open-source and shared-learning across scientific sub-domains. As such, we hope this award inspires all potential recipients to innovate, share and integrate (and repeat as needed).
# Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>5</td>
</tr>
<tr>
<td>Keynote Presentations</td>
<td>7</td>
</tr>
<tr>
<td>Abstracts table of contents</td>
<td>8</td>
</tr>
<tr>
<td>Abstracts</td>
<td>15</td>
</tr>
<tr>
<td>List of participants</td>
<td>110</td>
</tr>
</tbody>
</table>
# Program

**Day 1 (to be announced)**

<table>
<thead>
<tr>
<th>Time</th>
<th>ID</th>
<th>Title</th>
<th>Author</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td></td>
<td>Opening remarks, Rafael Najmanovich &amp; Ilan Samish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:05</td>
<td>K1</td>
<td>Communicating and Interacting with the Molecular Cell</td>
<td>Arthur Olson &amp; David Goodsell, Scripps, USA</td>
<td></td>
</tr>
<tr>
<td>10:05</td>
<td>20</td>
<td>The quixotic quest for perfection at high resolution</td>
<td>Daniel Keedy, Duke, USA</td>
<td></td>
</tr>
<tr>
<td>10:25</td>
<td></td>
<td>Coffee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td>K2</td>
<td>Modeling RNA Structure and Dynamics</td>
<td>François Major, U. Montreal, Canada</td>
<td></td>
</tr>
<tr>
<td>11:30</td>
<td>12</td>
<td>Genome-wide analysis of DNA shape provides new insights into protein-DNA recognition</td>
<td>Remo Rohs, USC, USA</td>
<td></td>
</tr>
<tr>
<td>11:50</td>
<td>35</td>
<td>The central role of post-translational modification in the interactome</td>
<td>Mark Wass, CNIO, Spain</td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td>Lunch &amp; Poster/Laptop session (odd ID numbers)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>K3</td>
<td>Decoding genetic variation to compute 3D structures of proteins</td>
<td>Chris Sander, MSKCC, USA</td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td>24</td>
<td>Identify druggable targets using microenvironments matching</td>
<td>Tianyun Liu, Stanford, USA</td>
<td></td>
</tr>
<tr>
<td>15:50</td>
<td>56</td>
<td>Predicting kinks in trans-membrane helices</td>
<td>Henry Wilman, Oxford, UK</td>
<td></td>
</tr>
<tr>
<td>16:10</td>
<td>21</td>
<td>A new structure and sequence dependent normal mode analysis model</td>
<td>Vincent Frappier, U. Sherbrooke, Canada</td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td></td>
<td>Coffee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:00</td>
<td>K4</td>
<td>The action of genotypes on phenotypes: less is more</td>
<td>Olivier Lichtarge, BCM, USA</td>
<td></td>
</tr>
<tr>
<td>17:30</td>
<td>13</td>
<td>The next generation of SCOP and ASTRAL</td>
<td>John-Marc Chandonia, LBNL, USA</td>
<td></td>
</tr>
<tr>
<td>17:50</td>
<td>50</td>
<td>Evolution of functions in domain structures - Exploring shifts in functional sites</td>
<td>Christine Orengo, UCL, UK</td>
<td></td>
</tr>
<tr>
<td>18:45</td>
<td></td>
<td>Pre-dinner reception (Location)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:30</td>
<td></td>
<td>Dinner (Location)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20:45</td>
<td></td>
<td>The Warren DeLano Structural Bioinformatics and Computational Biophysics Award</td>
<td>Margaret DeLano</td>
<td></td>
</tr>
<tr>
<td>20:55</td>
<td></td>
<td>Introduction to main 3Dsig keynote, Nobel Laureate Ada Yonath (by Ilan Samish)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21:00</td>
<td>K5</td>
<td>What was first? The genetic code or its products?</td>
<td>Nobel Laureate Ada Yonath WIS, Israel</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>ID</td>
<td>Title</td>
<td>Author</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>9:00</td>
<td>K6</td>
<td><strong>Mapping the protein universe with (mostly) structural genomics</strong></td>
<td>Adam Godzik</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SBMRI, USA</td>
<td></td>
</tr>
<tr>
<td>9:30</td>
<td>88</td>
<td>Exploring the evolution of protein function in Archaea</td>
<td>Alexander Goncearencu, U. Bergen, Norway</td>
<td></td>
</tr>
<tr>
<td>9:50</td>
<td>52</td>
<td>Format déjà vu: PDBX/MMCIF, the new data format for the WWPDB</td>
<td>John Westbrook, Rutgers, USA</td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td>38</td>
<td>Protein structure alignment beyond spatial proximity</td>
<td>Sheng Wang, TTIC, USA</td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td>Coffee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td>7</td>
<td>Bayesian integration of physics and knowledge-based potential functions</td>
<td>Martin Mechelke, Max Planck Institute for Developmental Biology, Germany</td>
<td></td>
</tr>
<tr>
<td>11:20</td>
<td>31</td>
<td>A probabilistic fragment based protein structure prediction algorithm</td>
<td>Kam Y. J. Zhang, Riken, Japan</td>
<td></td>
</tr>
<tr>
<td>11:40</td>
<td>59</td>
<td>Threading multiple subunits into an EM density map: application to the 26S proteasome</td>
<td>Keren Lasker, UCSF, USA</td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td></td>
<td>Lunch &amp; Poster/Laptop session (Even ID numbers)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>32</td>
<td>Organizing molecular principles of the molecular chaperone dynamics and regulatory interactions</td>
<td>Gennady Verkhivker, Chapman U., USA</td>
<td></td>
</tr>
<tr>
<td>15:20</td>
<td>58</td>
<td>Multi-conformer contact network analysis provides the structural and functional basis for NMR relaxation experiments</td>
<td>Henry van den Bedem, Stanford, USA</td>
<td></td>
</tr>
<tr>
<td>15:40</td>
<td>47</td>
<td>Conformational diversity on mono and multiple-domain proteins: relationship with protein evolution</td>
<td>Gustavo Parisi, U. Quilmes, Argentina</td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>D1</td>
<td><strong>Discussion:</strong> Protein structure and Function in light of Evolution</td>
<td>Olivier Litcharge, Adam Godzik, Christine Orenge, Chris Sander &amp; Rafael Najmanovich</td>
<td></td>
</tr>
<tr>
<td>16:40</td>
<td></td>
<td>Coffee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:10</td>
<td>19</td>
<td>Molecular docking with ligand &amp; target flexibility in FlexAID: Application to virtual screening</td>
<td>Francis Gaudreault, U. Sherbrooke, Canada</td>
<td></td>
</tr>
<tr>
<td>17:30</td>
<td>30</td>
<td>Structure-based ligand discovery for solute carrier transporters</td>
<td>Avner Schlessinger, UCSF, USA</td>
<td></td>
</tr>
<tr>
<td>17:50</td>
<td>2</td>
<td>Biomolecular motors and switches: from machines to drugs</td>
<td>Barry J. Grant, U. Michigan, USA</td>
<td></td>
</tr>
<tr>
<td>18:10</td>
<td>K7</td>
<td><strong>Characterizing the 3D environment of drug binding sites for personalized medicine</strong></td>
<td>Russ Altman, Stanford</td>
<td></td>
</tr>
<tr>
<td>18:40</td>
<td></td>
<td>Closing remarks, Philip Bourne</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:45</td>
<td></td>
<td>End of 3DSIG 2012</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Keynote Presentations

Keynote 1 - 13th @ 9h05
Communicating and Interacting with the Molecular Cell
Arthur Olson & David Goodsell, Scripps, USA

Keynote 2 - 13th @ 11h00
Modeling RNA Structure and Dynamics
François Major, U. Montreal, Canada

Keynote 3 - 13th @ 15h00
Decoding genetic variation to compute 3D structures of proteins
Chris Sander
MSKCC, USA

Keynote 4 - 13th @ 17h00
The action of genotypes on phenotypes: less is more
Olivier Lichtarge, BCM, USA

Keynote 5 - 13th @ 21h00
What was first? The genetic code or its products?
Nobel Laureate Ada Yonath
WIS, Israel

Keynote 6 - 14th @ 9h00
Mapping the protein universe with (mostly) structural genomics
Adam Godzik
SBMRI, USA

Keynote 7 - 14th @ 18h10
Characterizing the 3D environment of drug binding sites for personalized medicine
Russ Altman, Stanford
## Abstracts table of contents

<table>
<thead>
<tr>
<th>Id</th>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Satnam Surae, Damien Farrell, Catherine Godson, Jens Nielsen &amp; Finian Martin</td>
<td>COMPUTATIONAL ANALYSIS OF BMP-RECEPTOR AND BMP-ANTAGONIST INTERACTIONS WITH A VIEW TO DESIGNING BMP SUPER-AgonISTS AND DOMINANT NEGATIVES</td>
</tr>
<tr>
<td>02</td>
<td>Guido Scarabelli &amp; Barry J. Grant</td>
<td>BIOMOLECULAR MOTORS AND SWITCHES: FROM MACHINES TO DRUGS</td>
</tr>
<tr>
<td>03</td>
<td>Amitava Roy &amp; Carol B. Post</td>
<td>LONG-DISTANCE CORRELATIONS OF RHINOVIRUS CAPSID DYNAMICS CONTRIBUTE TO UNCOATING AND ANTIVIRAL ACTIVITY</td>
</tr>
<tr>
<td>04</td>
<td>Constance Jeffery</td>
<td>MOONLIGHTING PROTEINS</td>
</tr>
<tr>
<td>05</td>
<td>Noah Daniels, Raghavendra Hosur, Bonnie Berger &amp; Lenore Cowen</td>
<td>REMOTE HOMOLOGY DETECTION OF BETA-STRUCTURAL MOTIFS USING SIMPLIFIED RANDOM FIELDS AND SIMULATED EVOLUTION</td>
</tr>
<tr>
<td>06</td>
<td>Inbal Sela-Culang, Shahar Alon, Anat Burkovitz &amp; Yanay Ofran</td>
<td>A SYSTEMATIC COMPARISON OF FREE AND BOUND ANTIBODIES PROVIDES EVIDENCE FOR ALLOSTERIC EFFECTS IN ANTIGEN BINDING</td>
</tr>
<tr>
<td>07</td>
<td>Martin Mechelke &amp; Michael Habeck</td>
<td>BAYESIAN INTEGRATION OF PHYSICS AND KNOWLEDGE-BASED POTENTIAL FUNCTIONS</td>
</tr>
<tr>
<td>08</td>
<td>Jose M. Duarte, Adam Srebnik, Martin A. Schärer &amp; Guido Capitani</td>
<td>EVOLUTIONARY PROTEIN PROTEIN INTERFACE CLASSIFICATION: APPLICATIONS AND PERSPECTIVES</td>
</tr>
<tr>
<td>09</td>
<td>Shruthi Viswanath, DVS Ravikant &amp; Ron Elber</td>
<td>IMPROVING PROTEIN-PROTEIN DOCKING ALGORITHMS WITH SIDE CHAIN REFINEMENT AND ATOMIC POTENTIALS</td>
</tr>
<tr>
<td>10</td>
<td>Tim Wiegels, Marco Biasini &amp; Victor S. Lamzin</td>
<td>MORE COMPLETE PROTEIN MODELS IN MACROMOLECULAR CRYSTAL STRUCTURE DETERMINATION</td>
</tr>
<tr>
<td>11</td>
<td>Harianto Tjong, Ke Gong, Lin Chen &amp; Frank Alber</td>
<td>3D GENOME UNIVERSE OF THE BUDDING YEAST</td>
</tr>
<tr>
<td>12</td>
<td>Tianyin Zhou, Lin Yang, Iris, Dror, Ana Carolina Dantas Machado &amp; Remo Rohs</td>
<td>GENOME-WIDE ANALYSIS OF DNA SHAPE PROVIDES NEW INSIGHTS INTO PROTEIN-DNA RECOGNITION</td>
</tr>
<tr>
<td>13</td>
<td>John-Marc Chandonia</td>
<td>THE NEXT GENERATION OF SCOP AND ASTRAL</td>
</tr>
<tr>
<td>Id</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>----</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>14</td>
<td>George Nicola, Tiqing Liu, Linda Hwang &amp; Michael Gilson</td>
<td>BINDINGDB: A PROTEIN-LIGAND DATABASE FOR DRUG DISCOVERY</td>
</tr>
<tr>
<td>15</td>
<td>Sampada Bagai, Chongbo Sun &amp; Tian Tang</td>
<td>CALCULATING POTENTIAL OF MEAN FORCE OF POLYETHYLENIMINE MEDIATED DNA ATTRACTION USING WEIGHTED HISTOGRAM ANALYSIS METHOD</td>
</tr>
<tr>
<td>16</td>
<td>Matthieu Chartier &amp; Rafael Najmanovich</td>
<td>ISOMIF: A NEW TOOL FOR THE DETECTION OF MOLECULAR INTERACTION FIELD SIMILARITIES</td>
</tr>
<tr>
<td>17</td>
<td>Dominic Duchêne, Francis Gaudreault, Eloïc Colombo, Antoine Désilets, Richard Leduc, Éric Marsault &amp; Rafael Najmanovich</td>
<td>IN SILICO DEVELOPMENT OF A NEW SERIES OF MATRIPTASE INHIBITORS</td>
</tr>
<tr>
<td>18</td>
<td>Mario Pujato, Thomas MacCarthy, Avi Bergman &amp; Andras Fiser</td>
<td>THE UNDERLYING MOLECULAR AND NETWORK LEVEL MECHANISMS IN THE EVOLUTION OF ROBUSTNESS IN GENE REGULATORY NETWORKS</td>
</tr>
<tr>
<td>19</td>
<td>Francis Gaudreault, Dominic Duchêne &amp; Rafael Najmanovich</td>
<td>MOLECULAR DOCKING WITH LIGAND &amp; TARGET FLEXIBILITY IN FLEXAID: APPLICATIONS TO VIRTUAL SCREENING</td>
</tr>
<tr>
<td>20</td>
<td>Daniel Keedy, Michael Prisant, Lindsay Deis, David Richardson, &amp; Jane Richardson</td>
<td>THE QUIXOTIC QUEST FOR PERFECTION AT HIGH RESOLUTION</td>
</tr>
<tr>
<td>21</td>
<td>Vincent Frappier, Jean-Guy Lehoux, Pierre Lavigne &amp; Rafael Najmanovich</td>
<td>A NEW STRUCTURE AND SEQUENCE DEPENDENT NORMAL MODE ANALYSIS MODEL</td>
</tr>
<tr>
<td>22</td>
<td>Abdullah Kahraman, Franz Herzog, Alexander Leitner, Thomas Walzthöni, Lars Malmstroem &amp; Ruedi Aebersold</td>
<td>TOWARDS STRUCTURAL SYSTEMS BIOLOGY WITH CROSS-LINK DATA AND MOLECULAR MODELING</td>
</tr>
<tr>
<td>23</td>
<td>Thomas Haschka, Catherine Etchebest, Eric Henon, Laurent Martiny &amp; Manuel Dauchez</td>
<td>THE DEPLETION OF CALCIUM IONS EXPLORED BY MD SIMULATIONS – MULASE STUDY ON THROMBOSPONDIN</td>
</tr>
<tr>
<td>24</td>
<td>Tianyun Liu &amp; Russ B Altman</td>
<td>IDENTIFY DRUGGABLE TARGETS USING MICROENVIRONMENTS MATCHING</td>
</tr>
<tr>
<td>25</td>
<td>B.A. van den Berg, M.J.T. Reinders, J.M. van den Laan, J.A. Roubos &amp; D. de Ridder</td>
<td>A RATIONAL PROTEIN REDESIGN METHOD FOR IMPROVED SECRETION YIELDS</td>
</tr>
<tr>
<td>26</td>
<td>Michal Jamroz, Andrzej Kolinski &amp; Daisuke Kihara</td>
<td>PREDICTION OF REAL-VALUE FLUCTUATION OF GLOBULAR PROTEINS</td>
</tr>
<tr>
<td>27</td>
<td>Edvin Fuglebak, Sandhya Tiwari, Siv M. Hollup, Tristan Cagnolini, Lars Skjaerven, Sven H. Grindhaug &amp; Nathalie Reuter</td>
<td>COMPARING NORMAL MODES OF PROTEIN STRUCTURES USING WEBnm@ 2.0</td>
</tr>
<tr>
<td>Id</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>----</td>
<td>------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>28</td>
<td>Shula Shazman, Jie Chen, Hunjoong Lee, Richard Mann &amp; Barry Honig</td>
<td>ON THE FLY – A DATABASE OF DROSOPHILA MELANOGASTER TRANSCRIPTION FACTORS (TF) AND TF BINDING SITES (TFBS)</td>
</tr>
<tr>
<td>29</td>
<td>Masahito Ohue, Yuri Matsuizaki, Nobuyuki Uchikoga, Takashi Ishida &amp; Yutaka Akiyama</td>
<td>MEGADOCK: A RAPID SCREENING SYSTEM OF PROTEIN-PROTEIN INTERACTIONS BY EXHAUSTIVE DOCKING</td>
</tr>
<tr>
<td>30</td>
<td>Avner Schlessinger &amp; Andrej Sali</td>
<td>STRUCTURE-BASED LIGAND DISCOVERY FOR SOLUTE CARRIER TRANSPORTERS</td>
</tr>
<tr>
<td>31</td>
<td>David Simoncini, Francois Berenger, Rojan Shrestha &amp; Kam Y. J. Zhang</td>
<td>A PROBABILISTIC FRAGMENT-BASED PROTEIN STRUCTURE PREDICTION ALGORITHM</td>
</tr>
<tr>
<td>32</td>
<td>Gennady Verkhivker, Nathan Lawless, Elizabeth Berrigan, Kristin Blacklock</td>
<td>ORGANIZING MOLECULAR PRINCIPLES OF THE MOLECULAR CHAPERONE DYNAMICS AND REGULATORY INTERACTIONS</td>
</tr>
<tr>
<td>33</td>
<td>Juan Esquivel-Rodriguez &amp; Daisuke Kihara</td>
<td>FITTING MULTIMERIC PROTEIN COMPLEXES INTO ELECTRON MICROSCOPY MAPS</td>
</tr>
<tr>
<td>34</td>
<td>Inbal Budowski-Tal, Yael Mandel-Gutfreund, &amp; Rachel Kolodny</td>
<td>PATCHBAG: A NOVEL APPROACH FOR EFFICIENT DETECTION OF PROTEIN STRUCTURAL SIMILARITY</td>
</tr>
<tr>
<td>35</td>
<td>Mark N Wass, Alfonso Valencia &amp; Bernard Jacq</td>
<td>THE CENTRAL ROLE OF POST TRANSLATIONAL MODIFICATION IN THE INTERACTOME</td>
</tr>
<tr>
<td>36</td>
<td>Bradley Hintze, Christopher Williams, Daniel Keedy, David Richardson &amp; Jane Richardson</td>
<td>IDENTIFICATION AND VISUALIZATION OF SYSTEMATIC ERRORS IN LOW-RESOLUTION PROTEIN MODELS</td>
</tr>
<tr>
<td>37</td>
<td>Anton I. Petrov, Ryan Rahrig, Blake Sweeney, Craig L. Zirbel &amp; Neodes B. Leontis</td>
<td>RNA 3D HUB, WEBFR3D, AND R3DALIGN - DATABASES AND WEB APPLICATIONS FOR RNA STRUCTURAL BIOINFORMATICS</td>
</tr>
<tr>
<td>38</td>
<td>Sheng Wang, Jianzhu Ma, Jian Peng &amp; Jinbo Xu</td>
<td>PROTEIN STRUCTURE ALIGNMENT BEYOND SPATIAL PROXIMITY</td>
</tr>
<tr>
<td>39</td>
<td>Hai Fang, David de Lima Morais &amp; Julian Gough</td>
<td>AN UPDATED COMPARISON OF METHODS FOR REMOTE HOMOLOGY DETECTION</td>
</tr>
<tr>
<td>40</td>
<td>Feng Zhao &amp; Jinbo Xu</td>
<td>A POSITION-SPECIFIC DISTANCE-DEPENDENT STATISTICAL POTENTIAL FOR PROTEIN STRUCTURE AND FUNCTIONAL STUDY</td>
</tr>
<tr>
<td>41</td>
<td>Abul H.J. Ullah &amp; Jay Shockey</td>
<td>THREE DIMENSIONAL STRUCTURE ELUCIDATION OF ACYL-COA-BINDING PROTEINS FROM TUNG OIL AND ARABIDOPSIS THALIANA USING ROBETTA FULL-CHAIN STRUCTURE PREDICTION SERVER</td>
</tr>
<tr>
<td>Id</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>----</td>
<td>-------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>42</td>
<td>Karine Audouze, Tudor Oprea, Søren Brunak &amp; Olivier Taboureau</td>
<td>FROM CHEMICAL TO SYSTEMS BIOLOGY: HOW NETWORK PHARMACOLOGY CAN CONTRIBUTE TO IT?</td>
</tr>
<tr>
<td>43</td>
<td>Jianzhu Ma, Jian Peng, Sheng Wang &amp; Jinbo Xu</td>
<td>PROTEIN THREADING WITH SPARSE SEQUENCE PROFILES</td>
</tr>
<tr>
<td>44</td>
<td>Stephen MacKinnon, Michael Garton Anatoly Malevanets &amp; Shoshana J Wodak</td>
<td>SYSTEMATIC SURVEY OF INTERTWINED HOMOMERS</td>
</tr>
<tr>
<td>45</td>
<td>Stefan Senn, Paul G. Falkowski, Yana Bromberg</td>
<td>STRUCTURAL RELATIONSHIPS OF OXIDOREDUCTASES</td>
</tr>
<tr>
<td>46</td>
<td>Juergen Haas, Tobias Schmidt, Tiziano Gallo Cassarino, Valentina Romano, Lorenza Bordoli &amp; Torsten Schwede</td>
<td>CAMEO - CONTINUOUS AUTOMATED EVALUATION OF PROTEIN STRUCTURE PREDICTION SERVERS</td>
</tr>
<tr>
<td>47</td>
<td>Diego Javier Zea, Maria Silvina Fornasari, Cristina Marino Buslje &amp; Gustavo Parisi</td>
<td>CONFORMATIONAL DIVERSITY IN MONO AND MULTIPLE-DOMAIN PROTEINS: RELATIONSHIP WITH PROTEIN EVOLUTION</td>
</tr>
<tr>
<td>48</td>
<td>Dietlind L. Gerloff, Edward L. Liaw, Eli Draizen, Felicia D. Kemp &amp; D. Steven Hall</td>
<td>AN EVOLUTIONARY LINK BETWEEN TWO APICOMPLEXAN PROTEIN FAMILIES WITH DISTINCT DISULFIDE BONDS</td>
</tr>
<tr>
<td>49</td>
<td>Izabella A.P. Neshich, Fábio de Moraes, José A. Salim, Ivan Mazoni, José G. Jardine &amp; Goran Neshich</td>
<td>SIZE MATTERS: SURFACE HYDROPHOBICITY INDEX (SHI) DESCRIBES THE IMPACT OF THE SIZE OF INTERFACE AREA ON OLMOMERIZATION DRIVEN BY HYDROPHOBIC EFFECT</td>
</tr>
<tr>
<td>50</td>
<td>Benoît Dessailly, Alison Cuff, Natalie Dawson, Robert Rentzsch, Christine Orengo</td>
<td>EVOLUTION OF FUNCTIONS IN DOMAIN STRUCTURES—EXPLORING SHIFTS IN FUNCTIONAL SITES</td>
</tr>
<tr>
<td>51</td>
<td>Dorota Latek, Pawel Pasznik, Slawomir Filipek</td>
<td>AN INTERACTIVE ONLINE PLATFORM FOR STRUCTURE MODELING OF G-PROTEIN COUPLED RECEPTORS</td>
</tr>
<tr>
<td>52</td>
<td>John Westbrook, Helen M. Berman, Jasmine Young &amp; Gerard J. Kleywegt</td>
<td>FORMAT DÉJÀ VU: PDBX/MMCIF, THE NEW DATA FORMAT FOR THE WWPDB</td>
</tr>
<tr>
<td>53</td>
<td>Yana Gofman, Simona Shats, Bernard Attali, Turkan Haliloglu &amp; Nir Ben-Tal</td>
<td>HOW DOES KCNE1 REGULATE THE KV7.1 POTASSIUM CHANNEL? MODEL STRUCTURE, MUTATIONS AND DYNAMICS OF THE KV7.1-KCNE1 COMPLEX</td>
</tr>
<tr>
<td>54</td>
<td>Yoonjoo Choi, Sumeet Agarwal &amp; Charlotte M. Deane</td>
<td>HOW LONG IS A PIECE OF LOOP?</td>
</tr>
<tr>
<td>55</td>
<td>Syed Hussain Basha, K Naresh Kumar, T Deepthi &amp; R Nalini Prasad</td>
<td>gB-gH-gL COMPLEX, A NOVEL ANTI HERPES DRUG TARGET ELUCIDATED IN MOLECULAR DOCKING STUDIES</td>
</tr>
<tr>
<td>Id</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>----</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>56</td>
<td>Henry Wilman, Dr. Charlotte Deane &amp; Dr. Jiye Shi</td>
<td>PREDICTING KINKS IN TRANS-MEMBRANE HELICES</td>
</tr>
<tr>
<td>57</td>
<td>Jamie Hill &amp; Charlotte M. Deane</td>
<td>FACTORS AFFECTING MEMBRANE PROTEIN THREADING</td>
</tr>
<tr>
<td>58</td>
<td>Henry van den Bedem, Gira Bhabha, Kun Yang, Peter Wright &amp; James Fraser</td>
<td>MULTI-CONFORMER CONTACT NETWORK ANALYSIS PROVIDES THE STRUCTURAL AND FUNCTIONAL BASIS FOR NMR RELAXATION EXPERIMENTS</td>
</tr>
<tr>
<td>59</td>
<td>Keren Lasker, Daniel Russel, Jeremy Phillips, Friedrich Foerster</td>
<td>THREADING MULTIPLE SUBUNITS INTO AN EM DENSITY MAP: APPLICATION TO THE 26S PROTEASOME</td>
</tr>
<tr>
<td>60</td>
<td>Sandor Vajda, Tanggis Bohnuud, Dmitri Beglov, David R. Hall, Maxim D. Frank-Kamenetskii1 &amp; Dima Kozakov</td>
<td>COMPUTATIONAL SOLVENT MAPPING REVEALS DRAMATIC EFFECT OF DNA BREATHING ON DUPLEX DNA REACTIVITY</td>
</tr>
<tr>
<td>61</td>
<td>Arunima Singh, Deirdre Coombe &amp; Robert J. Woods</td>
<td>COMPUTATIONAL ANALYSIS OF GLYCOSAMINOGLYCAN BINDING TO CHEMOKINE CCL5</td>
</tr>
<tr>
<td>62</td>
<td>Micah Steffek, Michael Elliot, Karen Stanger, Rami Hannously, Tony Giannetti, Barthelemy Demeule, Christine Pozniak, Joe Lewcock &amp; Jeremy Murray</td>
<td>RESOLVING THE PARADOX OF CASPASE-6 SELF-ACTIVATION: CRYSTALLOGRAPHIC &amp; BIOPHYSICAL INSIGHTS INTO CASPASE-6 REGULATION</td>
</tr>
<tr>
<td>63</td>
<td>Shruti Kouli, Archana Achalere, Neeru Sharma, Uddhavesh Sonavane &amp; Rajendra Joshi</td>
<td>MOLECULAR DYNAMICS STUDIES TO PROBE THE LOSS IN THE DNA BINDING PROPERTY OF P53 HOT SPOT MUTATIONS</td>
</tr>
<tr>
<td>64</td>
<td>Chin-Hsien Tai, Dukka KC &amp; Byungkook Lee</td>
<td>INTERNALLY SYMMETRIC PROTEINS WITH MULTIPLE AXES</td>
</tr>
<tr>
<td>65</td>
<td>Federica Chiappori, Ivan Merelli, Luciano Milanesi &amp; Anna Marabotti</td>
<td>THE ANALYSIS OF INTERACTIONS BETWEEN GALK ENZYME AND KNOWN INHIBITORS SUGGESTS GUIDELINES TO DEVELOP NEW DRUGS FOR GALACTOSEMIC PATIENTS</td>
</tr>
<tr>
<td>66</td>
<td>Nils J. D. Drechsel, César L. Ávila, Michael A. Johnston, Jordi Villà-Freixa</td>
<td>DEVELOPMENT OF A MULTISCALE FRAMEWORK TO STUDY THE ENERGETICS OF PROTEIN DYNAMICS IN PROTEIN-PROTEIN INTERACTIONS</td>
</tr>
<tr>
<td>67</td>
<td>Anna Marabotti, Angelo Facchiano, Luciano Milanesi, Manshu Tang &amp; Kent Lai</td>
<td>STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF MUTANTS OF GALT ENZYME</td>
</tr>
<tr>
<td>68</td>
<td>Andrew J. Bordner, Barry Zorman &amp; Ruben Abagyan</td>
<td>EFFICIENT MODELING OF PEPTIDE FOLDING, ORIENTATION, AND ASSEMBLY IN LIPID BILAYERS USING IMPLICIT SOLVATION</td>
</tr>
<tr>
<td>Id</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>----</td>
<td>-------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>69</td>
<td>Sikander Hayat &amp; Arne Elofsson</td>
<td>SELECTING THREE-DIMENSIONAL TRANSMEMBRANE BETA BARREL MODELS USING PREDICTED Z-COORDINATES</td>
</tr>
<tr>
<td>70</td>
<td>Yuri Matsuzaki, Nobuyuki Uchikoga, Masahito Ohue &amp; Yutaka Akiyama</td>
<td>APPLICATION OF EXHAUSTIVE PROTEIN-PROTEIN INTERACTION PREDICTION SYSTEM BY USING PROTEIN DOCKING TO SIGNAL TRANSDUCTION PATHWAYS</td>
</tr>
<tr>
<td>71</td>
<td>Anna vangone, Daniele Sblattero, Ivana Caputo, Luigi Cavallo &amp; Romina Oliva</td>
<td>PREDICTION AND ANALYSIS OF IDIOTYPE-ANTI-IDIOTYPE ANTIBODY COMPLEXES ASSOCIATED TO CELIAC DISEASE</td>
</tr>
<tr>
<td>72</td>
<td>Lazaros Mavridis, Vishwesh Venkatraman, &amp; David W. Ritchie</td>
<td>A COMPREHENSIVE COMPARISON OF PROTEIN STRUCTURAL ALIGNMENT ALGORITHMS</td>
</tr>
<tr>
<td>73</td>
<td>Angelo Facchiano, &amp; Anna Marabotti</td>
<td>A SEMI-AUTOMATIC PROCEDURE FOR THE COMPARISON OF CONFORMATIONAL AND FUNCTIONAL PROPERTIES OF PROTEIN MUTANTS</td>
</tr>
<tr>
<td>74</td>
<td>Alexander Kantardjiev</td>
<td>QUANTUM ENTANGLED STATES IN PROTEIN-LIGAND DOCKING QUANTUM.LIGAND.DOCK-OPENCL/GPU IMPLEMENTATION</td>
</tr>
<tr>
<td>75</td>
<td>Marcin J. Skwark &amp; Arne Elofsson</td>
<td>ULTRA RAPID, ACCURATE QUALITY ASSESSMENT OF PROTEIN STRUCTURE MODELS</td>
</tr>
<tr>
<td>76</td>
<td>Timothy Nugent &amp; David T. Jones</td>
<td>ACCURATE DE NOVO STRUCTURE PREDICTION OF LARGE TRANSMEMBRANE PROTEIN DOMAINS USING FRAGMENT-ASSEMBLY AND CORRELATED MUTATION ANALYSIS</td>
</tr>
<tr>
<td>78</td>
<td>Sebastian Kelm, Charlotte M. Deane, Frank von Delft</td>
<td>FRAGMENT-BASED LOOP MODELLING FOR X-RAY CRYSTALLOGRAPHY OF MEMBRANE PROTEINS</td>
</tr>
<tr>
<td>79</td>
<td>Hannah Edwards &amp; Charlotte M. Deane</td>
<td>STRUCTURAL PROPERTIES OF PROTEINS AND THEIR RELATIONSHIPS WITH FOLD AGE</td>
</tr>
<tr>
<td>80</td>
<td>Yang Shen &amp; Bruce Tidor</td>
<td>TARGETING DRUG-RESISTANT HIV-1 PROTEASE MUTANTS USING INHERENT PROTEIN FLEXIBILITY</td>
</tr>
<tr>
<td>81</td>
<td>Yosi Shibberu &amp; Mark E. Brandt</td>
<td>CONTACT GEOMETRY OF ESTROGEN RECEPTOR DYNAMICS</td>
</tr>
<tr>
<td>82</td>
<td>Markus Wiederstein, Markus Gruber, Karl Frank &amp; Manfred J. Sippl</td>
<td>EXPLORING UNCHARTED TERRITORY IN PROTEIN ASSEMBLY SPACE</td>
</tr>
<tr>
<td>Id</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>----</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>83</td>
<td>Robert J. Fieldhouse, Amanda Rochon &amp; A. Rod Merrill</td>
<td>THE TOXIN ADP-RIBOSYLTRANSFERASES: DISCOVERY, STRUCTURAL MODELLING AND VALIDATION</td>
</tr>
<tr>
<td>84</td>
<td>Susan T. Mashiyama, Kyriacos Kouparris, Conor R. Caffrey, James H. McKerrow &amp; Patricia C. Babbitt</td>
<td>A GLOBAL COMPARISON OF SEQUENCES AND STRUCTURES OF THE HUMAN AND T. BRUCEI DEGRADOMES GIVES INSIGHTS INTO POSSIBLE PARASITE DRUG TARGETS</td>
</tr>
<tr>
<td>85</td>
<td>Irene Farabella, Gabriel Waksman &amp; Maya Topf</td>
<td>ALLOSTERIC COMMUNICATIONS IN THE OUTER-MEMBRANE TRANSLOCATION DOMAIN OF THE PAPC USHER PROTEIN?</td>
</tr>
<tr>
<td>86</td>
<td>J. Dongun Kim, Stefan Senn, Arik Harel, David A. Case, Vikas Nanda &amp; Paul G. Falkowski</td>
<td>EVOLUTION OF PROTEINS STUDIED BY SECONDARY STRUCTURE COMPOSITION</td>
</tr>
<tr>
<td>87</td>
<td>Paula M. Petrone</td>
<td>RETHINKING MOLECULAR SIMILARITY: COMPARING COMPOUNDS BASED ON BIOACTIVITY</td>
</tr>
<tr>
<td>88</td>
<td>Alexander Goncarenco &amp; Igor N Berezovsky</td>
<td>EXPLORING THE EVOLUTION OF PROTEIN FUNCTION IN ARCHAEA</td>
</tr>
<tr>
<td>89</td>
<td>Jesse Eickholt &amp; Jianlin Cheng</td>
<td>PREDICTING RESIDUE-RESIDUE CONTACTS AND DISORDERED REGIONS WITH DEEP NETWORKS</td>
</tr>
<tr>
<td>90</td>
<td>Swati Jain, Laura Murray, Jane S Richardson, Bruce R Donald &amp; David C Richardson</td>
<td>MODELING RNA X ANGLE FLEXIBILITY IN A SOFTWARE FOR STRUCTURE-BASED RATIONAL DESIGN OF RNA AND RNP COMPLEXES</td>
</tr>
<tr>
<td>91</td>
<td>Jorge LS Pina, Beatrix SLP Lima, Solange Guimarães &amp; Floriano P Silva Junior</td>
<td>DEVELOPMENT OF A PARTICLE SWARM OPTIMIZATION VARIANT OF AUTODOCK FOR EFFICIENT AND ACCURATE DOCKING OF HIGHLY FLEXIBLE LIGANDS</td>
</tr>
<tr>
<td>92</td>
<td>Todd J. Taylor &amp; Iosif Vaisman</td>
<td>STATISTICS OF THE GEOMETRY AND TOPOLOGY OF REAL AND MODEL PROTEIN STRUCTURES</td>
</tr>
<tr>
<td>93</td>
<td>Elchin S. Jufayev, Ryan J. McLaughlin, Yi-Ping Tao &amp; William A. McLaughlin</td>
<td>KB-ROLE: AN ONLINE RESOURCE FOR THE IDENTIFICATION OF PREDICTED FUNCTIONS OF PROTEIN STRUCTURES BASED ON CHAIN LEVEL AND RESIDUE LEVEL ANNOTATIONS</td>
</tr>
<tr>
<td>94</td>
<td>Chaim A. Schramm, Brett T. Hannigan, Jason E. Donald, Chen Keasar, Jeffrey G. Saven, William F. DeGrado &amp; Ilan Samish</td>
<td>KNOWLEDGE-BASED POTENTIAL FOR POSITIONING MEMBRANE ASSOCIATED STRUCTURES AND ASSESSING RESIDUE SPECIFIC ENERGETIC CONTRIBUTIONS FOR PREDICTION AND DESIGN</td>
</tr>
</tbody>
</table>
Abstracts
INTRODUCTION

Bone Morphogenetic Proteins are integral components of bone and organ development. BMPs are secreted extracellular proteins and stimulate gene expression by associating with membrane bound receptors. BMPs are closely regulated by a family of protein antagonists, including Gremlin, Noggin and Crossveinless-2, which modulate BMP-receptor association by occupying and thereby occluding receptor binding of BMP. Gene expression of greml1 is upregulated and correlates to the progression of diabetic nephropathy. Hence, we have analysed the BMP-receptor and BMP-antagonist interactions using a newly designed automated pipeline, Protein Complex Tool (PCT), which utilises PEAT-SA. Co-crystal structures of the protein-protein interactions were submitted to PCT which performed computational alanine substitution scans to calculate the free energy contribution of each BMP-2,-7 residue to the stability of the complexes with receptors, BMPRIs (BMP2-7) and ActRIIAs (BMP2-7), and antagonists, Crossveinless-2 (BMP2-7) and Noggin (BMP-7 only). Saturation scans were also performed on BMP-2,-7 to identify potential specificity switching mutations.

METHODS

We analysed the BMP-receptor and BMP-antagonist interactions using a newly designed automated pipeline, Protein Complex Tool (PCT). PCT utilises PEAT-SA and allows for the rapid and easy prediction of Δ∆Gfold values for each residue in a protein-protein interaction. PCT displays information about the PDB file, controls the extraction of the relevant chains from the PDB file, automatically invokes PEAT-SA, calculates Δ∆Gfold and displays the results in an automatically generated bar plot and a mapped protein structure.

Stability calculations are performed by PEATSA which estimates changes in the protein free energy using different energy contributions, such as Van der Waals, electrostatic, hydrogen bond, desolvation energies and entropy terms. The Δ∆Gfold is the difference between the free energy values of the wild-type and mutant structures. The stability calculations are performed for each mutation, for each protein chain and for the complex. The Δ∆Gfold is the difference in Δ∆Gfold of the complex and the constituents of the complex. For example, we carried out this analysis on a BMP dimer (B) in complex (B+P) with a binding partner dimer (P), either a receptor or antagonist, which is described by the following equation:

\[ Δ\DeltaG_{\text{fold}}(B+P) = Δ\DeltaG_{\text{fold}}(B) + Δ\DeltaG_{\text{fold}}(P) - Δ\DeltaG_{\text{fold}}(B+P) \]

RESULTS & CONCLUSIONS

Co-crystal structures of the protein-protein interactions were submitted to PCT and computational alanine substitution scans were performed to calculate the free energy contribution of each BMP-2,-7 residue to the stability of the complexes with receptors, BMPRIs (BMP2-7) and ActRIIAs (BMP2-7), and antagonists, Crossveinless-2 (BMP2-7) and Noggin (BMP-7 only). The values from all five scenarios were summarised to a heat map (Fig. 1) where the colouring represents the relative destabilising (red) or stabilising (blue) effects of the mutation on complex stability. The free energy calculations by alanine scan identified the ‘wrist’ and ‘knuckle’ binding epitopes and also showed conservation of binding between BMP-2 and BMP-7.

Further mutational analysis was performed using saturation scans – mutating each residue to the other 19 amino acids. We identified superagonists for both BMP-2: L51V and N102T and BMP-7: E60T, D119I, I124A and K127E. We also identified dominant negatives for both BMP-2: S88G and L92D and BMP-7: F117E and V122D. The super-agonists will bind and activate receptor but will be resistant to binding by antagonist; in contrast, the dominant negatives, bind antagonist but not receptor. These identified mutations have a specificity switch value in the same range as those calculated for an experimentally validated super-agonist, E60K. These mutations will be tested experimentally to evaluate their effects on the receptor and antagonist interactions. We have identified a short-list of potential therapeutic protectants derived from BMPs by designing an automated pipeline which integrates the steps for structure based energy calculations.

REFERENCES

BIOMOLECULAR Motors AND SWITCHES: FROM MACHINES TO DRUGS

Guido Scarabelli* & Barry J. Grant†

*Department of Computational Medicine and Bioinformatics, University of Michigan. †bjgrant@umich.edu

Molecular motors and switches lie at the heart of key biological processes, from the division and growth of cells to the muscular movement of organisms. Our approach to studying these fascinating nanomachines couples bioinformatics (to probe sequence-structure-function relationships); molecular dynamics (to investigate essential conformational changes); Brownian dynamics (for diffusional protein-protein encounters); and computer-aided drug design (for discovering novel therapeutics). I will describe two discoveries that exemplify the power of this approach. First, how it uncovered the importance of electrostatics in the motion of kinesin motors, and how this information enabled the rational design of mutant motors with tailored velocities. Second, how it revealed that the traditional “induced fit” view for activating conformational changes in molecular switches should be replaced by a “conformational selection” model, and how this framework led to the discovery of novel small molecule Ras inhibitors.

INTRODUCTION

Cytoskeletal molecular motors and related G-protein molecular switches are responsible for powering and regulating diverse cellular functions including intracellular transport and signal transduction (Figure 1). They represent attractive targets for drug design, as their aberrant function is associated with many diseases including cancer, diabetes and neurodegenerative disorders.

![Figure 1.](image)

**Figure 1.** We are interested in deciphering the mechanisms by which molecular motors and switches function and how their dysfunction is related to disease.

METHODS

To understand how these fascinating nanomachines function requires the consideration of multiple spatial and temporal scales as well as the successful integration of experiment, molecular simulation and theory. We have developed a state-of-the-art multi-level computational approach to investigate the structure, dynamics and interactions of prototypical motor and switch systems. Our approach couples bioinformatics (to probe sequence-structure-function relationships); molecular dynamics (to investigate essential conformational changes); Brownian dynamics (for diffusional protein-protein and protein-ligand encounters); and computer-aided drug design (for discovering novel therapeutics).

RESULTS & CONCLUSIONS

Comparative experimental structure analysis coupled with accelerated molecular dynamics simulations revealed the intrinsic ability of the molecular switches Ras and Rho to interconvert between active and inactive conformations in the absence of nucleotide1-3 (Figure 2A). These results support an underlying conformational selection mechanism and also highlighted the conserved dynamic coupling of the nucleotide-binding site with distal transient small molecule binding pockets4-5. Candidate binders for these novel pockets were selected through ensemble docking of small molecule compound libraries6. Finally, cell-based assays confirmed our hypothesis that the chosen binders can inhibit the downstream signaling activity of Ras7. We thus propose that the predicted sites are viable targets for the development and optimization of new allosteric drugs.

We have recently found, using Brownian dynamics simulations combined with experimental mutagenesis, that incoming kinesin motor heads undergo electrostatically-guided diffusion-to-capture by microtubules, and that this produces enhanced directionally-biased binding8 (Figure 2B). Real-world mutagenesis of individual residues, predicted to be important for electrostatic guidance powerfully influences kinesin-driven microtubule sliding, with one mutant producing a 6-fold acceleration over wild type9. We conclude that electrostatic interactions play an important role in the kinesin stepping mechanism, by biasing the diffusional association of kinesin with microtubules, and that rationally engineered charge substitutions can be used to produce faster velocity kinesin motors.

![Figure 2.](image)

**Figure 2.** (A) Distinct conformations of Ras from accelerated molecular dynamics simulations. (B) Sampling density of kinesin about tubulin from Brownian dynamics simulation highlight preferred association paths. Further images and animations related to this work can be found at: [http://foggrantlab.org](http://foggrantlab.org)

REFERENCES

INTRODUCTION

Human rhinovirus (HRV) and other members of the enterovirus genus bind small-molecule antiviral compounds in a cavity buried within the viral capsid protein VP1 [1]. These compounds block the release of the viral protein VP4 and RNA from inside the capsid during the uncoupling process [2, 3, 4, 5, 6] and prevent “breathing” motions, the transient externalization of the N-terminal regions of VP1 and VP4 from the inside of intact viral capsid [7]. We conduct molecular dynamics (MD) simulations to explore how the antiviral compound, WIN 52084, alters properties of the HRV 14 capsid through long-distance effect. We developed an approach to analyze capsid dynamics in terms of correlated radial motion and the shortest-paths [8] in the network of these motions. We observe [9] in the absence of WIN, correlated radial motion is observed between residues separated by as much as 85 Å. The most frequently populated path segments of the network were localized near the 5-fold symmetry axis and included those connecting the N-termini of VP1 and VP4 with other regions, in particular near 2-fold symmetry axes, of the capsid. The results provide evidence that the virus capsid exhibits concerted long-range dynamics, which have not been previously recognized. Moreover, the presence of WIN destroys this radial correlation network, suggesting that the underlying motions contribute to a mechanistic basis for the initial steps of VP1 and VP4 externalization and uncoating.

METHODS

We calculated trajectories of either HRV 14 (Tnowin) and HRV 14 with WIN 52084 (Twin) bound using rotational boundary symmetry conditions [10] for a total time equal to 30 ns each. We calculated radial correlation, CR(i, j), Pearson’s correlation coefficient between magnitude of vectors, values for C α atoms of non-interacting residue pairs, i.e. pairs with a minimum distance between residues greater than the non-bond cutoff distance 14 Å. To trace long distance radial correlation, we developed a method for detecting the link between two C α atoms based on a modified version of Dijkstra’s graph searching algorithm [8]. The nodes of the network correspond to C α atoms and edges connecting all C α atoms belonging to residues within the non-bonded distance of 14 Å. An edge between C α i and C α j was assigned a value equal to 1 – CR(i, j). The path between any two nodes i and j of the network, with n ≥ 1 intermediate nodes in between, was determined from the minimum value of the weight

\[ W_{i_1...i_n} = 1 - CR(i, j) \cdot CR(i, j) \cdot ... \cdot CR(i, j) \cdot CR(i, j) \]

The betweenness of an edge is the fraction of all possible shortest paths that include the edge.

RESULTS & CONCLUSIONS

Many of the Tnowin pairs with radial correlation >0.6 were separated by distances >45 Å and some with distances as large as 85 to 90 Å. This was not the case for Twin; many fewer C α – C α pairs separated by long distances have a 95% confidence interval for CR(i, j) greater than 0.6. The radially correlated C α – C α pairs separated by distances greater than 45 Å are largely non-overlapping sets for Tnowin and Twin. To trace the long distance radial correlation, we calculated shortest path between every possible C α – C α pairs and calculated betweenness for every C α – C α pairs. Numerous path segments are observed from Tnowin and found to be concentrated around the canyon and the 5-fold axis. The set of lines encircling the 5-fold axis corresponds to the N-terminal region of VP3, which forms a well-ordered β-annulus through inter-protoemer interactions. This annulus is absent from the pattern of highest betweenness paths calculated from Twin when WIN is bound. Residues present in these pathway are mostly conserved in HRVA and HRV-B families. Residues present in ten unique edges with highest betweenness contains known spontaneous drug resistant residues. Together, the pattern of radial correlations and highest-betweenness paths comprised of conserved residues, and the loss of this pattern when WIN is bound strongly suggest these paths are the origins of the radial correlations and part of the functional motion leading to uncoating.

REFERENCES

MOONLIGHTING PROTEINS

Constance Jeffery
Dept. Biological Sciences, University of Illinois at Chicago. cjeffery@uic.edu

The number of known "moonlighting proteins" is continuing to increase, with over 150 identified to date. My lab is collecting information about the sequences, structures, and functions for the known moonlighting proteins for the purpose of constructing a database of moonlighting proteins.

INTRODUCTION

Moonlighting proteins comprise a class of multifunctional proteins in which a single polypeptide chain has multiple biochemical functions that are not due to gene fusions, splice variants or multiple proteolytic fragments. Many of the known moonlighting proteins are cytosolic enzymes that are also transcription factors, crystallins, chaperones, extracellular growth factors, or cell surface adhesins (1-8).

The diverse examples of moonlighting proteins already identified, the multiple methods by which one protein can switch between functions, the two proposed models for the evolution of moonlighting functions, and the potential benefits moonlighting proteins can provide through coordinating cellular activities, suggest that moonlighting proteins might be common. The ability of a single protein to have multiple functions and to be involved in different multi-protein complexes or biochemical pathways adds to the difficulty in predicting protein functions from sequences or structures and in interpreting the results of proteomics projects. The ability of proteins to have more than one function could also complicate database annotation, especially automated annotation based on sequence homology, because homologues of moonlighting proteins might have only one of the multiple functions.

To date, most moonlighting proteins have been found by serendipity. There is currently no straightforward method to identify proteins with moonlighting functions, or for determining if a protein of interest is a moonlighting protein. Information about moonlighting proteins is scattered in many publications by multiple labs with some small collections of information about the proteins and their functions contained in review articles. There is also some confusion in the literature because some authors have used the term “moonlighting” when the publication is actually describing a protein family in which each member has a different function, a protein that has the same function but in two different locations (two biological roles but not two biochemical functions), or a protein that is another type of multifunctional protein.

In this project, we are collecting information about the sequences, structures, and functions for the over 150 moonlighting proteins for which multiple biochemical functions have been experimentally verified. The long term goal is to include that information in a web-based centralized location, the MoonProt Database.

METHODS

We are collecting and organizing information about the known moonlighting proteins from the literature. Analysis of the sequences and structures is underway. The long term goal is for the information to be made available in a manually curated, searchable database of moonlighting proteins, the MoonProt Database. The database will include sequences, links to structures, descriptions of functions, locations of functional sites in sequences and on structures, locations of where the functions are exhibited in the cell or organism, and information about pathways. The MoonProt database will be similar in format to the DisProt database (9).

RESULTS & CONCLUSIONS

We have identified over 150 moonlighting proteins from the literature for which at least two biochemical functions have been experimentally verified.

The MoonProt database will help organize information about the known moonlighting proteins and provide a verified set of proteins for analysis. The database could serve as a standard for the future development of improved methods for predicting which proteins have multiple biochemical functions. Researchers in other fields will be able to look up the proteins they are studying and see if they are known to moonlight. This information could be useful in many types of research projects, for example, when interpreting data from systems biology or proteomics projects.

REFERENCES

We introduce a Markov Random Field-based approach, SMURFLite, for remote homology detection in beta-structural proteins. SMURFLite outperforms contemporary methods at beta-structural motif recognition, and remains computationally tractable on all beta-structural protein folds. We have applied it to whole-genome annotation on the Thermotoga Maritima genome.

Profile hidden Markov models (HMMs) have been among the most successful methods to date for recognizing protein sequences that are evolutionarily related\(^1\). However, these models do not capture pairwise (non-local) statistical preferences of residues that are hydrogen bonded in beta sheets. These dependencies have been partially captured in the HMM setting by simulated evolution\(^2\) and can be fully captured by Markov random fields (MRFs)\(^3\). However, the MRFs can be computationally prohibitive when beta strands are interleaved in complex topologies.

We present SMURFLite, a method that combines both simplified Markov random fields and simulated evolution to substantially improve remote homology detection for beta structures. Unlike previous MRF-based methods, SMURFLite is computationally feasible on any beta-structural motif.

Markov random fields have been previously suggested by many\(^4\) as a generalization of HMMs with more power for remote homology detection. However, when designing MRFs for this problem, two design difficulties emerge. First, the MRF becomes more difficult to train. Second, it quickly becomes computationally intractable to find the optimal-scoring parse of a target to the model.

We employ a simplified MRF, parameterized by an *interleave number*, representing the number of other beta strands that appear between a pair of hydrogen-bonded beta strands in sequence, plus one (Figure 1). Our dynamic-programming approach is exponential in this interleave number. To keep MRFs tractable, we simply remove those conditional probabilities (beta-strand interactions) that exceed a user-specified threshold. For example, in Figure 1, SMURFLite might remove the dependency between beta strands a and b, while preserving their roles in the HMM as well as preserving strand b’s dependency with strand c. This would reduce the exponent in the run-time from 4 to 2.

With no added query-time cost (and negligible training-time cost) we replace those graphical dependencies with additional, artificial training data\(^7\) that we generate based on those same conditional probabilities.

We test SMURFLite on all propeller and barrel folds in the mainly-beta class of the SCOP hierarchy in cross-validation experiments. We show a mean 26% (median 16%) improvement in AUC for beta-structural motif recognition as compared to HMMER, a mean 33% (median 19%) improvement over RAPTOR\(^8\), and even a mean 18% (median 10%) improvement over HHpred\(^9\), despite HHpred’s use of extensive additional training data. We demonstrate SMURFLite’s ability to scale to whole genomes by running a SMURFLite library of 207 beta-structural SCOP superfamilies against the entire genome of *Thermotoga maritima*, and make over a hundred new fold predictions.

![Figure 1](image)

**Figure 1.** A closed beta barrel (PDB ID 1bw3, a Barwin domain) from the superfamily “Barwin-like endoglucanases” to illustrate interleaving of strand pairs. Beta strands a and b, which close the barrel, have interleave 4, while strands c and d, which are adjacent in sequence, have interleave 1. Strands b and c have interleave 2.

A webserver that runs SMURFLite is available at: [http://smurf.cs.tufts.edu/smurflite/](http://smurf.cs.tufts.edu/smurflite/)

For more information, please see our paper:


**REFERENCES**

A SYSTEMATIC COMPARISON OF FREE AND BOUND ANTIBODIES PROVIDES EVIDENCE FOR ALLOSTERIC EFFECTS IN ANTIGEN BINDING

Inbal Sela-Culang1,2, Shahr Alon1, Anat Burkovitz2 & Yanay Oftran1
The Goodman Faculty of Life Sciences, Bar Ilan University1,2; inbal@ofranlab.org

We systematically compared the free and antigen-bound structures of 49 antibodies solved in these two forms. While there are hardly any systematic binding-related changes to the antigen binding site itself, there are binding-related conformational changes throughout the Fab, including the constant regions. These changes provide evidences for allosteric effects in the antibody that are related to antigen binding.

INTRODUCTION
It is widely accepted that antibody (Ab) specificity is determined solely by the variable region. Thus most of the structural changes upon antigen (Ag) binding are assumed to occur in the variable domains, and specifically in the six CDRs. However recent reports suggest that Ag binding, as well as the structure of the variable domain, may be correlated with changes in the constant domains1-5. To study the structural changes that occur in antibodies upon binding of an antigen, we systematically compared the free and antigen-bound structures of 49 antibodies that were experimentally solved in these two forms.

METHODS
We retrieved all Abs with experimentally determined 3D structures in both free and bound forms. To characterize the structural changes that occur in Abs upon Ag binding, we calculated for each residue in each Ab: (i) the RMSD between the free and the bound form, (ii) the change in solvent accessibility and (iii) the change in the number of inter-chain contacts. Results from structurally equivalent positions of different Abs were averaged over all Abs. When possible, the same calculation was also performed for pairs of bound structures of the same Ab in order to elucidate whether the changes are Ag-binding related. Inter domain changes were measured by superimposing two equivalent domains, and calculating the Cα RMSD between the two other domains. Analysis of somatic hypermutations was done on a separate data-set of Ab-Ag complexes, regardless of the existence of a free structure. The germline sequence of each Ab was identified and aligned to the mature sequence. The probability of mutations in each position was calculated based on this alignment.

RESULTS & CONCLUSIONS
We found that the Ag binding-site remains fairly rigid upon Ag binding, as CDR-H3 is the only CDR to show a significant conformational change (1.3Å averaged Cα RMSD) and even for this CDR, a substantial structural change of >1Å was observed only for 37% of the Abs. The region showing the most prominent and consistent conformational change between free and bound structures is the CH1-1 loop, located on the heavy chain constant domain, far from the Ag binding site (Figure 1). This loop is a part of the interface with the light constant domain, thus having a potential effect on the heavy-light relative orientation, which may be translated by the elbow angle to changes in the Ag binding-site (see below), or vice versa. The higher frequency of somatic hypermutations in the CH1-1 loop relative to other positions in the Fab constant domains, suggests that it may be related to Ag binding. In addition, this loop is intrinsically disordered in more than one third of the structures, implying that it may mediate binding to other proteins, yet to be discovered.

Structural changes upon Ag binding also occur in the relative orientation of the Fab domains. The relative orientation of the heavy-light chains changes more when comparing free-bound than when comparing bound-bound structures, in agreement with the idea that heavy-light relative orientation contributes to Ag binding. The relative orientation of the variable-constant domains undergoes a greater change than that of the heavy-light chains. The change observed in free-bound is greater than the change in bound-bound, indicating that it is related to Ag-binding. We propose a possible mechanism for the role of this flexibility of the elbow angle, in which Ag binding causes a change in the VH-VL relative orientation, which in turn, may change the contacts between residues in the variable-constant interface, resulting in an altered elbow angle. This may be translated into a different heavy-light relative orientation in the constant region, and may be possibly related to the binding of an effector to the constant domains. An additional support for the relationship between the change in the elbow angle and Ag binding is its proportion to the Ag size: larger Ags cause more substantial changes than those observed for smaller Ags. Our observations support the existence of allosteric effects in the Fab structure proposed recently.

REFERENCES

![Figure 1. The structure of BH-151 (PDB code 1eob), colored according to the average free-bound Ca, RMSD, from blue (low RMSD) to red (high RMSD). Regions which for the RMSD was not calculated are colored black.](image-url)
Bayesian Integration of Physics and Knowledge-Based Potential Functions

Martin Mechelke* and Michael Habeck

Max Planck Institute for Developmental Biology Tübingen, Germany, *martin.mechelke@tuebingen.mpg.de

Potential functions that judge if a conformation is energetically favorable are at the heart of many application in computational structural biology. Two different approaches to derive potentials are commonly used: physics-based approaches approximate the fundamental laws of physics, while knowledge-based approaches are built upon information extracted from databases of known protein structures. A fusion of these complementary approaches seems appealing, but raises the question of how to combine the potentials to avoid bias and double counting without losing information. We show a solution how a linear combination of physics and knowledge-based potentials can be adaptively calculated, based on a rigorous application of Bayesian model comparison. We apply our approach to infer the weight of a knowledge-based dihedral potential and show that weighted potential functions are unbiased and produce significantly more accurate protein structures.

INTRODUCTION

In protein structure calculation, all experimental data need to be interpreted in the light of our prior knowledge on biomolecular structures, usually encoded as a potential function. The choice of potential function can have substantial impact on the outcome of a structure calculation. Ideally the potential function would guide the calculation towards the native state without contradicting the experimental data. But how should we combine physical and knowledge based functions without overfitting or introducing bias?

METHODS

We use the experimental data to guide the selection of the energy functions. To keep the problem general and tractable we introduce a weight \( \beta \) on the knowledge-based energy function. Using a Bayesian approach the optimal \( \beta \) by calculating the marginal likelihood, which is the probability of observing the data for a particular choice of \( \beta \). We select the weight that maximises the marginal likelihood and thus is best supported by the experimental data. The marginal likelihood is defined as an integral over all degrees of freedom, the evaluation of which is highly challenging. An extended replica-exchange Monte Carlo sampling scheme [1] in combination with histogram reweighting techniques for estimation of the free energy allows us to solve the problem of estimating the marginal likelihood and generating structures from the posterior distribution simultaneously and efficiently.

RESULTS & CONCLUSIONS

To illustrate our approach, we introduce a knowledge based potential for backbone dihedral angles in structure calculation. First, we study the effect of Bayesian weighting on the posterior ensemble for high quality data available for ubiquitin (PDB id:1d3z). In this case the marginal likelihood is minimal for \( \beta = 0.97 \). Although 1d3z is a high-quality data set, our addition of a weighted knowledge-based potential still improved the structure quality as measured by Procheck and WhatCheck.

To further validate our method, we check how well the posterior ensembles for different weights are supported by residual dipolar coupling (RDC) and scalar couplings, which were not included in the calculation. In both cases Bayesian weighting achieves almost minimal errors (Figure 1).

To study how Bayesian weighting performs in more challenging structure determination problems, we choose a sparse data set for the Fyn-SH3 domain (1zbj) as well as a data set measured by solid-state NMR spectroscopy for the c-spectrin SH3 domain. As for ubiquitin, we observe a significant improvement in the quality of the generated ensemble as indicated by the Procheck and WhatCheck validation criteria.

As shown in Figure 1, the optimal weight is found in regions where the average RMSD to the crystal structure shows a minimum.

Conclusions. We outline a new formalism to integrate physics- and knowledge-based potentials based on data driven principles.

REFERENCES

EVALUATION PROTEIN PROTEIN INTERFACE
CLASSIFICATION: APPLICATIONS AND PERSPECTIVES
José M. Duarte1, Adam Srebnick2, Martin A. Schärer1 & Guido Capitani1*
1Biomolecular research, Paul Scherrer Institut, Switzerland. 2SyBio, ETH Zurich, Switzerland.

Distinguishing crystal contacts from biologically relevant interfaces is an important issue in protein structure analysis that naturally lends itself to evolutionary approaches. We present here a new method (and software) using three criteria, two of which evolutionary, to classify protein interfaces. The method compares favourably to PISA in performance and is suitable for a variety of applications in structural bioinformatics and structural biology, for instance validation of structures and homology models and divide-and-conquer approaches to the structure determination of supramolecular complexes. We have implemented the method as a command line tool and an easy-to-use web server.

INTRODUCTION

The problem of distinguishing crystal contacts from biologically relevant interfaces has emerged as an important challenge to experimental and computational biologists alike. Since biological interfaces are subject to selection pressure while crystal contacts are not, methods that detect that pressure are in principle the most logical ones to distinguish the two types of interfaces. We have recently described a proof-of-concept evolutionary method, CRK [1], using Ka/Ks ratios as metrics of selection pressure. We present here a new approach, greatly improving on CRK in recall, speed and applicability, which employs sequence entropy to measure selection pressure.

METHODS

Our method, called EPPIC (Evolutionary Protein Protein Interface Classification), uses three criteria to evaluate interfaces: two evolutionary and a geometric one. First, the ratio of average sequence entropies acting on interface core and rim residues is computed. Second, the average entropy of core residues is compared to that of randomly pooled surface residues. If an interface contains N core residues, 100 random pools of N surface residues are drawn and used to calculate a Z score-like value:

\[ Z = \frac{H_{core} - \mu_{random}}{\sigma_{output}} \]

Third, the number of interface core residues is used as geometric indicator of interface character. We use Schärer's [1] definition of core residues, i.e. those residues burying more than 95% of their ASA upon complexation. The three criteria are then combined by majority to produce a final call (“xtal” or “bio”) of the interface character.

RESULTS & CONCLUSIONS

Optimization datasets. We optimized parameters and thresholds for EPPIC versus two newly compiled datasets of large crystal contacts and small biological interfaces. The interface area distribution of our datasets extensively covers the “overlap range” where the two kinds of interfaces coexist.

Performance. We benchmarked EPPIC against a minimally updated version of the Ponsoingl [2] dataset of crystal and biological interfaces. Its performance, with an accuracy of 86%, compares favourably to that of PISA (84%), which is the current state-of-the-art method in the field and does not use evolutionary information [3].

Web interface. EPPIC has been implemented both as command-line tool and as a user-friendly web interface (Figure 1), which provides the user with a description and a call (xtal or bio) for all interfaces in the crystal.

Can it be applied to homology models? A question we are addressing is whether oligomeric or protein complex interfaces in homology models can be recognized by EPPIC as biological and if the method can be used for model quality control. The first evidence being collected is that, indeed, interfaces of good quality homology models are classified by EPPIC as biological. An example is provided by the homology model of the dimeric enzyme GAD65 [4], published two years before the crystal structure of the protein appeared [5]. This represents a particularly successful case of homology modelling (2.1 Å rmsd to the crystal structure on 773 aligned Cτatoms), based on two templates with 22% and 14% sequence identity. EPPIC (Table 1) classifies the dimer interface in the model as “bio”, exactly with the same pattern of calls as the interface in the crystal structure, albeit with a smaller number of core residues, partly due to the incomplete coverage of the model.

TABLE 1. EPPIC results for the crystal structure and model of GAD65

<table>
<thead>
<tr>
<th>File</th>
<th>Interf. area (Å²)</th>
<th>Core size</th>
<th>Core-rim entropy</th>
<th>CSZ</th>
<th>Final</th>
</tr>
</thead>
</table>

FIGURE 1. EPPIC web server results for PDB entry 2OKK[5]

Conclusions. EPPIC is a very effective protein interface classifier using evolutionary information. Among its most promising applications are structure and homology model validation and its use in correctly assembling components of large supramolecular entities, like the nuclear pore complex, in divide-and-conquer approaches [6].

REFERENCES

IMPROVING PROTEIN-PROTEIN DOCKING ALGORITHMS WITH SIDE CHAIN REFINEMENT AND ATOMIC POTENTIALS

Shruthi Viswanath1, DVS Ravikanth2 & Ron Elber3

Depts. of Computer Science1, Chemistry and Biochemistry1 and Institute of Computational Engineering and Science3, University of Texas at Austin, @WalmartLabs.2, ron@ices.utexas.edu

We propose an improvement to rigid protein-protein docking algorithms by incorporating side chain refinement and energy minimization, followed by reranking the refined models by a combination of residue and atomic contact potentials. Our algorithm is tested on three separate datasets, and consistently provides improvements over rigid docking. Our docking package is available at http://clsb.ices.utexas.edu/dock_by_pie

INTRODUCTION

Structural characterization of protein-protein interactions provides important insight into biological processes and protein function. Many docking algorithms that computationally predict the structure of complexes, sample exhaustively, all possible transformations of the ligand. Transformations are then scored using residue or atomic potentials. Several methods that refine and rerank the rigid docking models have been recently proposed. They include Monte-Carlo based minimization in RosettaDock1 and minimization followed by evaluation of electrostatic and desolvation energy in RDOCK2 among others.

METHODS

We use our previously developed docking algorithm, DOCK/PIE3 to generate the FFT-based transformations for a receptor-ligand pair. Transformations are scored based on a residue contact potential4. The transformations with a high number of inter-atomic clashes (more than 45 clashes) are discarded. The remaining transformations are clustered based on the interface RMSD and the first ten cluster centers of the transformations are designated as the output complexes from this rigid docking procedure.

In order to improve the contact energies of the docked decoys (models) and make the structures more chemically reasonable, i.e. reduce the number of clashes, we performed side chain refinement on the top one thousand models from the rigid docking procedure described in the previous paragraph. We chose top thousand models as it was observed that we could find near native models (models with interface RMSD less than 2.5 A) in the top thousand, in most of the targets in our datasets. We first modeled the side chains of the interface residues (residues of one protein which are within 6.5 A from the interacting protein) using rotamers. We used SCWRL4 for this purpose. We then performed a 100 step rigid body minimization in vacuum to further refine the models. We used the MD package, MOIL5 to do the minimization.

We then reranked the refined structures using an atomic contact potential, which has 32 coarse-grained heavy atom types and 3 distance bins, 2-3.5 A, 3.5-5 A and 5-8 A. The parameters of the atomic potential were determined by a mathematical programming approach, training on models from 640 complexes6. Inequalities were formulated, comparing the atomic contact energy of the native structure of a target to that of each of the top thousand refined models. Additional inequalities comparing the energy of near-native models (iRMSD to the native was less than 2.5 A) with that of misaligned models (iRMSD to the native greater than 7 A) were added. The linear programming package, PE3 was used to solve the resulting 5.8 million inequalities. A solution that satisfied 92.8 percent of the inequalities was obtained. The atomic potential was then applied to rerank the top thousand refined models, in order to enrich the number of hits in the final top ten models. We however, found that though the atomic potential uncovers more hits in the top 100 than the residue potential5, it is less sensitive than the residue potential when it comes to the top 10. The reason may be the application of the algorithm to approximate complexes (unbound) rather than to bound complexes. We used a combination of the residue and atomic potentials, taking the product of their scores, to rerank the models after refinement. The intuition is that models can be more accurately identified as hits, if they have low energies in both residue and atomic potentials.

RESULTS & CONCLUSIONS

We tested the refinement and atomic filter on three different test sets: a test set of 60 novel unbound-unbound targets that were separate from the learning set, and consisted of complexes deposited in the PDB after Sept 2010, 103 targets from the ZLAB Benchmark 3.0 set, and 579 targets from the learning set. Figure 1 shows the number of targets with hits (models with interface RMSD less than 4 A) in the top 10, for rigid docking (DOCK/PIE), and refinement plus reranking applied to rigid docking (RR-DOCK/PIE) for all three datasets.

REFERENCES

We present a novel method for improving model completeness and fragmentation of models automatically obtained from medium-to-low-resolution electron density maps in macromolecular crystal structure determination. We exploit the presence of non-crystallographic symmetry (NCS), as well as complementary information from the increasing abundance of structural information in the Protein Data Bank (PDB).

INTRODUCTION
Macromolecular machines play central roles in life processes and present important targets in biomedical and pharmaceutical research. However, determining the three-dimensional structures of large molecular assemblies is a challenging task in macromolecular crystallography (MX). Crystals of such structures rarely diffract to high resolution and often only noisy and inaccurate electron density maps are obtained. Computational approaches for model building in MX have historically been focused on high-resolution data, thus their application to data extending to lower than 3.0 Å resolution is limited and typically results in incomplete and highly fragmented models. Hence, methods to improve the completeness and the accuracy of models are urgently needed for the automated structure determination at low-resolution.

To address this aim within the ARP/wARP software project [1], we exploit the fact that 50% of all crystal structures deposited in the PDB [2] contain multiple copies of subunits or their assemblies in the asymmetric unit. These copies are related by the so-called non-crystallographic symmetry, NCS. We noticed that during automated model building with ARP/wARP, particularly in its initial steps, NCS-related parts of the structure are rarely built in exactly the same way. However, a beneficial side effect of differently built NCS-related copies is that each provides information that is not present in another copy; combining this (which we call intrinsic) information improves the model building process and increases the overall completeness of built structures, particularly at low resolution.

Density that cannot be easily interpreted as part of a protein chain can be regarded as a poorly defined connection between two built chain fragments. Such connection generally contains not only loops but also helices or strands. In an approach using complementary information, we use structural fragments from the PDB for the interpretation of such density. These approaches resulted in two complementary methods; the PNS-extender [3] – for model extension using NCS and FittOFF – for fitting PDB fragments into non-interpreted density (see figure 1).

METHODS
The PNS-extender follows an all-versus-all least squares superposition of intermediate models resulting from ARP/wARP model building. Matches are clustered according to their rotational difference and NCS relations are identified. For fragments from differently built NCS-related copies, found relations are used to extend and connect fragments in subsequent building cycles. FittOFF identifies areas of non-interpreted density between consecutive chain fragments. Statistical observations and secondary structure prediction are used to identify the length of a loop to be fitted between the fragments. Backbone conformations are sampled from a large fragment database created from high-resolution X-ray structures and scored by spatially correlating them to the residual density. Top-ranking fragment conformations from both methods are fed back to the ARP/wARP model building process as new seed points. The iterative nature of ARP/wARP helps to avoid model bias.

RESULTS & CONCLUSIONS
Testing the basic function of NCStry was conducted by deleting fragments worth 35% of structural information from a pentameric high-resolution structure (PDB Id 1c48). NCStry was used to retrieve the missing residues and the result was superposed to the reference structure. The complete structure was rebuilt with an RMSD of just 0.3 Å. The use of NCS during model building with ARP/wARP also provides a significant improvement in many cases, and often requires less model-building cycles. In the best case, at 3.2 Å resolution, the model completeness improves from 55% to 73% (27% more residues were built), more side chains are docked in sequence, and the length of the built fragments increases. Preliminary tests for FittOFF showed very promising results. Rebuilding of up to 6-residue long gaps in high-resolution structures can already be achieved with an RMSD of under 0.5 Å. Application to model building is currently being tested and will be presented at the conference.

**FIGURE 1.** Model completeness and fragmentation is improved by using NCS (intrinsic) information and fitting fragments from a structural database into not interpreted electron density (complementary).
Unlike a protein, the 3D folding of chromosomes will not yield a unique structure while performing numerous functions in nuclei. Genome structures are inherently plastic, with large cell-to-cell variations among cells of the same type and under the same conditions. To address the challenge of modeling highly variable genome structures, we propose a population-based modeling approach, where we construct a large population of 3D genome models. We define a scoring function as the sum of all spatial constraints that act together on all models in the population. We interpret the result in terms of probabilities of a sample drawn from a population of heterogeneous structures. We quantitatively characterize the emerging contact patterns between individual loci and their nuclear territories that arise when chromosomes are allowed to behave as constrained but otherwise randomly configured flexible polymer chains. Remarkably, such a constrained random encounter model is sufficient to explain in a statistical manner the experimentally determined hallmarks of the budding yeast genome organization. We show that tethering of heterochromatic regions to nuclear landmarks and random encounters of chromosome confined in the nucleus control the higher order organization of the yeast genome.

INTRODUCTION
We quantitatively analyze the organization of the budding yeast genome in light of the tethering of heterochromatic regions to nuclear landmarks in the context of a confining nuclear architecture, the competition between chromosomes for space, and the purely random encounters of chromosome chains. The population of genome structures that results from this constrained random encounter model agrees remarkably well with data from genome-wide conformation capture and fluorescence imaging experiments. With remarkable consistency, the population reproduces chromatin interaction frequencies, gene territories, the relative distances between telomeres, and even the spatial clustering of functionally related chromosome regions such as early replication start sites and tRNA gene loci. Our findings therefore indicate that the tethering of a few heterochromatic chromatin regions to nuclear landmarks and the competition between strands in a confined nuclear architecture are the major organizing principles of the *S. cerevisiae* genome. Specific interactions between chromatin regions, although possibly also present, are not required to explain the available experimental data on the higher-order genome organization of this species. Within a cell population of uniform genotype, individual genome structures can vary dramatically. Due to this structural variation, any investigation of genome structure-function relationships should employ population analysis, and interpret its results in terms of probabilities. The presentation of a single representative structure, or even an average genome structure, does not adequately reflect the wide range of structural features that can appear in the genome, nor can it explain the universe of functional associations that can arise in individual cells.

METHODS
To generate a population of three-dimensional genome structures, we defined an optimization problem with three main components: (1) a structural representation of chromosomes as chromatin fibers; (2) a scoring function quantifying the genome structure’s accordance with nuclear landmark constraints; and (3) an optimization method, which minimizes the scoring function to generate a population of genome structures that entirely satisfies all landmark constraints.

The optimization is performed using a combination of simulated annealing molecular dynamics and the conjugate gradient methods implemented in the Integrated Modeling Platform (IMP). An individual optimization starts with an entirely random bead configuration, followed by an initial optimization of the structure. Next, we apply simulated annealing protocols to entirely equilibrate the genome configuration. Finally, conjugate gradient optimization ensures that all constraints are satisfied, leading to a structure with score zero. Many independent optimizations are carried out to generate a population of 200,000 genome structures with a total score of zero, hence consistent with all input data (Figure 1).

RESULTS & CONCLUSIONS
Our results emphasize the considerable structural variability of genome structures in individual cells. Each property of the simulated 3D genome population is compared with available experimental data and show remarkable agreement. We therefore demonstrate that most aspects of genome organization can be explained without calling on biochemically mediated chromatin interactions. The fact that geometrical constraints alone lead to a highly organized genome structure on which different functional elements are specifically distributed has strong implications for the folding of the genome structure and the evolution of its function.

REFERENCES
The sequence-dependent shape of DNA, such as the geometry of the major and minor groove, plays an important role in mediating interactions between protein and DNA molecules that are critical for carrying out cellular functions. Whereas high-throughput sequencing technologies continue to produce large amounts of DNA sequence information, experimental data on DNA three-dimensional structures are limited. To bridge this gap, we have developed a high-throughput method for predicting shape of naked DNAs on a genomic scale.

RESULTS & CONCLUSIONS

We used our new high-throughput approach to analyze the DNA shape of thousands of Hox protein binding sites derived from SELEX experiments combined with massively deep sequencing. We predicted the shape of more than 650,000 sites in total for all eight *Drosophila* Hox proteins, including two isoforms of Ultrabithorax (Figure 1A). We built a Euclidean distance tree solely based on DNA shape and found the eight Hox proteins ordered according to their collinearity (Figure 1B). This result revealed that anterior and posterior Hox proteins select for different DNA shape (Figure 1C). Previous research has demonstrated that the location of Hox genes at the chromosome is collinear with their expression profile along the anterior-posterior axis of the fly embryo. To these properties, we now add DNA shape as another characteristic that distinguishes anterior from posterior Hox genes. Knowledge about DNA shape provides, thus, additional clues how Hox genes have differentiated throughout evolution. We have also used this approach for analyzing a large number of nucleosome binding sites and have shown that a previously known periodic distribution of sequence elements [2] also leads to a periodicity of the three-dimensional structure of DNAs not bound by histones.

**REFERENCES**

THE NEXT GENERATION OF SCOP AND ASTRAL

John-Marc Chandonia

Physical Biosciences Division, Lawrence Berkeley National Laboratory, JMChandonia@lbl.gov

We released new versions of both SCOP and ASTRAL (1.75A) in March 2012. The new releases are the first to be fully dependent on a new SQL-based infrastructure and build procedure, and the first to be presented to the public through a single, unified interface. They also represent the first public deployment of our fully automated classification scheme: more than 11,000 new PDB entries were added to the current release, without sacrificing the reliability that SCOP has accumulated through years of careful manual curation. We plan to introduce additional features in a series of stable releases, while a major reclassification (SCOP 2.0) is in progress.

INTRODUCTION

The Structural Classification of Proteins (SCOP) database is a manually curated, near-comprehensive ordering of domains from proteins of known structure in a hierarchy according to their structural and evolutionary relationships. The ASTRAL compendium is a collection of software and databases, closely related to SCOP, that is used to aid research into protein structure and evolution. We are currently redesigning both databases to reflect recent advances in both biology and computer technology, as well as introducing more automated curation in order to stay up to date with the increasing number of protein structures. We recently released new versions of both databases under a single, unified interface, and we are continuing to improve the interface in order to meet the current needs of our users.

METHODS

We benchmarked an automated classification algorithm (ASTEROIDS) against all stable versions of SCOP from 1.55 to 1.75, in order to identify the most reliable subset of predictions. To do this, we built ASTEROIDS against each version of SCOP and used it to classify all new PDB entries that were added to the next SCOP release. We checked the predicted domains for all superfamilies that were not reclassified (e.g., merged or split) by manual curation. A predicted domain was considered to be correct if it was in the correct superfamily, and had no significant differences from the manually curated domain boundaries.

We found that predictions that met a set of fairly conservative criteria (domains spanning an entire PDB chain, which matched another undivided PDB chain from one of the first 7 SCOP classes, and were not annotated as low-resolution or synthetic) actually had an error rate lower than that of manual curation. Manual examination of all differences between these predictions and manually curated domains (150 differences out of 19,310 total predictions) revealed 2 errors in domains that had been manually curated in prior versions of SCOP, as well as several truncation errors in domains that had been automatically classified in prior versions of SCOP, and no case in which the prediction was clearly less accurate than the manually curated domain. We also noticed a number of previously automatically classified domains that were classified incorrectly at the Species or Protein levels. All such errors were corrected by manual curation in SCOP 1.75A.

To fully classify these predictions in the SCOP hierarchy, we also had to assign levels below Superfamily. Based on our benchmarking, we developed heuristic rules to classify domains at the Family, Protein, and Species levels. In cases where the protein or family could not be reliably matched to an existing SCOP clade, we created a new clade called "automated matches" rather than risking inaccurate predictions.

RESULTS & CONCLUSIONS

Using the new automation procedure, we added 11,152 PDB files to SCOP 1.75A that were not classified in SCOP 1.75. We also replaced all old automatically assigned domains in SCOP, in cases where an updated prediction was available. This resulted in the re-annotation of an additional 10,421 domains.

SCOP and ASTRAL 1.75A feature a new SQL-based back end as well as a single, unified web interface. Many objects that were difficult to find in the original layout, such as the change history, are now available under tabs. Thumbnail images were automatically generated to show each SCOP domain on its own and in several structural contexts, and these are displayed as part of the browser.

PLANS

We are continuing to refine our automated classification process, allowing more PDB entries to be classified automatically without sacrificing accuracy. Once we are satisfied that the automation is sufficiently robust, we will release another stable release that includes more new PDB files, and also introduce weekly automated updates. Because many users rely on SCOP (and the ASTRAL subsets) to benchmark algorithms, we will continue to periodically release manually curated stable releases, while also augmenting these stable releases with weekly updates that include automated classifications of new PDB entries.

Our Cambridge team, led by Alexey Murzin, has created a working prototype of SCOP 2.0, a major reclassification of the SCOP database. The prototype has limited but representative content, and is being readied for a public test release. SCOP 2.0 differs from its predecessor in two fundamental aspects. First, the tree-like hierarchy has been replaced with a complex network of hierarchical and non-hierarchical relationships represented by a directed acyclic graph, which allows curators to represent structural and evolutionary relationships between proteins in more precise detail. Second, there is no division of proteins into domains with fixed boundaries; instead, node-specific domain boundaries define the protein regions for each particular relationship.

REFERENCES


The large and growing body of experimental data on biomolecular binding is of enormous value in developing a deeper understanding of molecular biology, in developing new therapeutics, and in various molecular design applications. However, most of these data are found only in the published literature and are therefore difficult to access and use. BindingDB is a public web-accessible database of measured binding affinities for various molecular types. The BindingDB allows queries based upon a range of criteria, including chemical similarity or substructure, sequence homology, numerical criteria and reactant names. The data specification includes significant experimental detail. The time and expense required to extracting data from the literature, for this and many other databases, highlight the importance of moving toward machine-readable components of publications.

INTRODUCTION

The large and growing body of experimental data on biomolecular binding is of enormous value in developing a deeper understanding of molecular biology, in developing new therapeutics, and in various molecular design applications. However, most of these data are found only in the published literature and are therefore difficult to access and use. BindingDB is a public web-accessible database of measured binding affinities for various molecular types. The BindingDB allows queries based upon a range of criteria, including chemical similarity or substructure, sequence homology, numerical criteria and reactant names. The data specification includes significant experimental detail.

FIGURE 1. BindingDB contains 800,000 binding data, for 6,000 protein targets and 350,000 small molecules.

UTILITIES AND FEATURES

Chemical Similarity

BindingDB provides users the ability to run a chemical similarity search with one of three methods:
1. Maximum similarity
2. Binary Kernel Discrimination (BKD)
3. Support Vector Machine (SVM)

Linking Affinity and Structure

Each affinity data is showcased in a separate page, and bidirectional links to the PDB are shown if available.

REFERENCES

A series of umbrella sampling MD simulations were performed to calculate the potential of mean force of attraction between two DNA molecules in the presence of polyethyleneimine (PEI) molecules. The effects of DNA/PEI charge ratio and protonation state of PEI were addressed.

**INTRODUCTION**

The phenomena of compacting DNA chains into dense particles, which can fit into a volume of radius $1\mu$m or less, is known as DNA condensation. DNA condensation has been of interest to scientists in various fields as it not only explains how genetic information is packed but also has potential applications for non-viral gene delivery. Experimentally, it has been shown that DNA condensation can occur in the presence of multivalent and polycationic ions. In particular, PEI has been identified as one of the effective polymers that can form toroidal nanoscale aggregates of DNAs to facilitate their cellular uptake and effective transfection. To quantitatively describe the attraction between DNAs aggregated by PEI, in this work, we performed molecular dynamics (MD) simulations and used the weighted histogram analysis method (WHAM) to calculate the potential of mean force (PMF) of the interaction between two DNAs bridged by PEI molecules. PEIs have a wide range of molecular weights, protonation ratios and degrees of branching that can affect the transfection efficiency of PEI-delivered genetic materials. The ratio of the charges on the DNA to that on the PEI (N/P charge ratio) is also an important factor affecting DNA condensation. Influences of the protonation state of PEI and N/P charge ratio on the PMF will be addressed in this work.

**METHODS**

The Dickerson-Drew B-DNA dodecamer used in the simulation, $d$CGCGAATTCGCG$\text{d}$, consists of 24 nucleotides and carries a total charge of $-22$ in its fully deprotonated state. The PEI used in the simulation has 13 amine groups and a molecular weight of 586 Da. A total of 5 systems were simulated. Each of the first four systems contains two DNAs and a different number (8, 6, 4 and 2) of 46% protonated PEIs. They are referred to as 2D-8P, 2D-6P, 2D-4P and 2D-2P systems respectively. The last system contains two DNAs and eight 23% protonated PEI and is named as 2D-8P(23%) system. The simulations were run using NAMD$^2$ with CHARMM general force field$^3$ for PEI and CHARMM 27 force field$^4$ for the rest of the molecules. Each system was placed in a water box large enough so that any DNA or PEI molecule would not interact with their nearest periodic images. Counterions were added to neutralize the systems. The distance between the centers of mass of the two DNAs was chosen as the reaction coordinate and WHAM was used to obtain the PMF. Umbrella sampling windows were chosen at 1 Å interval along the reaction coordinate varying from 22 Å to 50 Å; and each window was run for 20 ns.

**RESULTS & CONCLUSIONS**

The PMF for the 2D-8P system is plotted along the reaction coordinate (Figure 1). The minimum value of the PMF is $-6.7$ kcal/mol obtained at 23 Å. The location of minimum is in accordance with the shortest distance between two DNAs aggregated by eight PEI molecules in a recent simulation performed by Sun et al.$^{11}$

**REFERENCES**

2. Wilson, R. W. and Bloomfield, V. A. Biochemistry 18, 2192-2196 (1979).
The detection of molecular interaction field similarities allow the identification of polypharmacological targets, can help prevent cross-reactivity, help guide the rational design of specific inhibitors and predict function. The program presented here, IsoMIF, allows the identification of molecular interaction field similarities across proteins, RNA and small molecules. The technique used can be applied to structurally distant molecules and does not require the superimposition of the molecules under comparison.

**INTRODUCTION**

There is a wide variety of methods able to detect similarities between molecules (proteins, RNA or small molecules) at the level of sequence, structure and molecular interaction fields. The latter can be used for the purpose of rational drug design, prevention of cross-reactivity, identification of polypharmacological targets and prediction of function. These methods require a set of probes and the definition of a potential energy function which are used to calculate the molecular interaction fields in a volume defined at the surface of the molecule of interest. Current methods (i.e.: 3DQSAR) that calculate and compare molecular interaction fields are limited regarding their fields of application as they require that the studied molecules be reliably superimposed. The current project aims to develop a method that allows the detection of molecular interaction fields similarities at the surface of proteins, RNA or around small molecules regardless of the structural similarities of the molecules under comparison.

**METHODS**

IsoMIF (Molecular Interaction Fields) works in two steps: (1) the calculation of potential energy vectors and (2) the identification of similar energy vectors positioned in geometrically equivalent positions in space. A grid with a resolution of 0.25 Å is built in the volume of a binding cavity or around a small molecule. At each grid intersection, a potential energy vector (composed of multiple potential interaction energies, each one calculated by a specific probe) is calculated. The interaction energy of a given probe is calculated with a potential energy function that contains two terms: a Lennard-Jones potential to approximate short-range interactions and a second term that accounts for long-range electrostatic Coulomb interactions. The probes and parameters of the energy functions corresponding to each probe or atom is defined by the force field used (e.g., MMF94, AMBER94, OPLS, CHARMM). Once the energy vectors are calculated at each grid intersection for a set of proteins, RNA or small molecules, a graph-matching algorithm is used to detect similar vectors that are in geometrically equivalent positions in space between the candidate molecules. The regions where similarities are found can be visualized using PyMOL (Schrödinger, LLC).

**RESULTS & CONCLUSIONS**

The effect of changing different parameters inherent to the IsoMIF program as well as those related to the probes and force field are currently being studied. For instance, we can take two identical protein structures that differ only by a single mutation and test the sensitivity of the program by looking at how the mutated residue affects similarities in its immediate vicinity as a function of the type of mutation. We are also evaluating how protein conformational changes influence the identification of molecular interaction fields similarities. Notwithstanding the need to optimize the different parameters, we can already study real cases (Figure 1). More precisely, we wish to study cases of single mutations in the vicinity of ligands with or without an effect on binding in terms of their effect on MIF similarities.

The method makes it possible to identify similarities between structurally distant molecules. For instance, different ATP binding proteins are recognized as drug targets for many diseases but can exhibit a large variety of folds and modes of ATP binding. As all ATP binding proteins are potential targets of ATP competitive inhibitors, IsoMIF could be used to detect similarities between proteins with different folds. IsoMIF provides an alternative to current techniques where the identification of molecular interaction field similarities across protein folds (or chemical scaffolds) is not possible. The similarities detected by IsoMIF between ATP binding proteins can help the rational design of inhibitors specific to subsets of ATP binding proteins. The information provided by IsoMIF can also be exploited to minimize cross-reactivity. The field of application of IsoMIF extends from rational drug design to the identification of polypharmacological targets, the prevention of cross-reactivity, the prediction of function and the study of protein-protein interactions.

**REFERENCES**

IN SILICO DEVELOPMENT OF A NEW SERIES OF MATRIPTASE INHIBITORS

Dominic Duchêne1, Francis Gaudreault1, Éloïc Colombo, Antoine Désilets2, Richard Leduc2, Éric Marsault2, Rafael Najmanovich3

1 Depts. of Biochemistry and Pharmacology, University of Sherbrooke; 2 Rafael.Najmanovich@USherbrooke.ca

INTRODUCTION

Type II transmembrane serine proteases (TTSPs) are a relatively new class of proteolytic enzymes with crucial roles in numerous physiological processes. The structure of TTSPs is composed of a cytoplasmic amino-terminal region, a transmembrane domain, a stem region composed of various functional domains and a carboxy-terminal serine protease catalytic domain. The catalytic domain is made of 8 subpockets corresponding to the 8 amino acids of the cleaved sequence. These amino acids are defined as P4-P3-P2-P1/P1’-P2’-P3’-P4’ where the cleavage site is between the P1 and P1’ residues. TTSPs are divided into four subfamilies: HAT/DESC, Hepsin/TMPRSS, Corin and Matriptase. Matriptase is one of the most studied members of TTSPs. Various studies have shown that its overexpression leads to various forms of epithelial cancer. Its implication in cancer makes it an interesting pharmacological target and one of the most studied members of TTSPs. Furthermore, it was shown that other members of TTSPs such as matriptase-2 and hepsin are implicated in tissue development and homeostasis. All of them have a highly conserved catalytic serine protease domain making the development of selective inhibitors difficult. The study of structural differences within the binding-sites of the different TTSPs can be exploited to create selective or multifunctional inhibitors. Based on our previously reported selective matriptase peptidomimetic inhibitor in the form of RQAR (see figure 1), we developed a new series of matriptase inhibitors.

METHODS

High-throughput docking studies were conducted to determine unique structural determinants to members of the TTSPs including matriptase, matriptase-2 and hepsin.

Molecular docking simulations were carried out with our in-house developed program Flexdocking® using our grid computing infrastructure (NRG@Home) with partial ligand (key dihedral bonds) and binding-site (specific side-chains) flexibility. Docking poses were re-scored with MM/GBSA (molecular mechanics generalized Born surface area) to find the best docking poses and to estimate binding free energies.

We docked an ensemble of 8000 peptidomimetic ligands structurally similar to natural substrates against the 3 targets. All ligands are composed of 1) four amino acids covering regions from P1 to P4, with an arginine at position P1 and 2) a serine trap in the form of a ketobenzothiazole group in P1’. The serine trap was selected to form a covalent but reversible bond with the catalytic serine. Selectivity profiles against all targets were calculated using the whole ensemble of peptidomimetic ligands in order to determine characteristics of matriptase that permit this selectivity against matriptase-2 and hepsin. Since the structure of matriptase-2 has not yet been solved experimentally, a homology model of the catalytic domain has been modeled with MODELLER 9v8 using matriptase as a template.

RESULTS & CONCLUSIONS

Firstly, we found some major structural differences between matriptase, matriptase-2 and hepsin in the S2-S3-S4 binding sub-pockets. Preferences for each binding sub-pocket were defined which led to the development and synthesis of new compounds. Such compounds are currently being tested in vitro and compared to the original RQAR peptidomimetic inhibitor. Here, we report a new series of selective inhibitors of matriptase developed by large scale virtual screening. Furthermore, the results permit us to understand the preferred types of interactions necessary in each sub-pocket to obtain selective matriptase inhibitors vis-à-vis matriptase-2 and hepsin. These preferences will be used to guide the creation of a library of non-peptidic compounds that will be used for large scale virtual screening and in vitro testing of these best scoring compounds.

REFERENCES

Robustness was observed as an emergent property of gene network evolution but the underlying molecular mechanisms are poorly understood. To explore these mechanisms we used a multi-tier gene network model integrating molecular sequence and structure information with network architecture and population dynamics. Results suggest that sparse architectures exploit a mixture of sequence and network level changes, whereas highly connected architectures evolve robustness entirely via network level changes.

**INTRODUCTION**

Robustness to genetic and environmental perturbations is ubiquitous in biological systems [1]. Previous studies using gene regulatory network models have shown how robustness evolves under conditions of stabilizing selection [2]. In this class of models, mutations alter only the interaction strengths, constraining the model to fixed network architectures. Therefore these models cannot explore potentially important factors such as evolved redundancy [3], modularity [4] and degeneracy [5]. The actual mechanisms underlying changes in network interactions will predominantly involve mutations at the sequence level in cis-regulatory regions [6] rather than changes in protein sequence and structure [7]. Even though DNA binding motifs of many transcription factors have been characterized [8], there is still a limited understanding of the evolutionary forces involved in the creation and maintenance of these motifs in the context of a gene regulatory network.

**METHODS**

The evolution of robustness in gene regulatory networks is likely to include mechanisms at both the sequence and network levels. Until now these two levels have been considered separately. Here we present a model that combines these two levels, enabling us to address their relative influence and how they interact in the context of the evolution of robustness (Fig. 1). In our model the gene regulatory network (Fig. 1B) is determined by transcription factor binding sites within DNA sequences, which undergo mutation. The effect of mutations on the binding affinity is estimated through the explicit modeling of DNA-protein complexes (Fig. 1A). Finally, the previous two levels are embedded in a population dynamics model to explore fitness effects (Fig. 1C). We categorize the types of mutations in a continuum ranging from silent mutations, which have no effect on regulation and change only the DNA sequence (sequence level), to mutations that change connections between genes in the network (network level).

**RESULTS & CONCLUSIONS**

We find that in sparse networks, a balance of sequence and network level mechanisms are responsible for the evolution of robustness, but when the network is densely connected the network-architecture level mechanisms become dominant. At the sequence level, robustness evolves by decreasing the probabilities of both the destruction of existent and generation of new binding sites. Meanwhile, in highly interconnected architectures, robustness evolves almost entirely via network level changes, deleting and creating binding sites that modify the network architecture. We argue that the shift towards the network level for more densely-connected networks offers a potential explanation for the evolution of increased complexity.

**REFERENCES**

Molecular Docking with Ligand & Target Flexibility

Francis Gaudreault1, Dominic Duchêne1 & Rafael Najmanovich2,

Department of Biochemistry, Faculty of Medicine and Health Sciences1, Université de Sherbrooke. Rafael.Najmanovich@USherbrooke.ca

We present FlexAID: a small-molecule docking algorithm that includes ligand and target flexibility (both small and large motions). FlexAID can successfully predict the complex structure of a large fraction of an unbiased dataset of 157 proteins cases despite the increasing number of variables introduced by flexibility. We discuss the use of FlexAID in virtual screening studies in the context of the Germination Protease of C. difficile.

Introduction
Small-molecule docking is used to predict the bound structure of a ligand and a target as well as predict function. The technique is mainly applied to virtual high-throughput screening studies (vHTS), due to its relatively short computational runtime. However, certain constraints are imposed, particularly to decrease the conformational search space. For example, the target molecule is most often treated as rigid. Previously, we demonstrated that introducing flexibility in the target is crucial1. We compared the bound and unbound structures of a large ensemble of non-redundant proteins and showed that side-chain movements are critical in approximately 30% of all binding-sites (excluding proteins undergoing large conformational changes). For these cases, rigid docking would fail to predict the right conformation of the bound complex. For cases in which docking succeeds in predicting the right conformation, most algorithms struggle when predicting the corresponding binding affinities. Thus, another critical issue in those studies is to accurately estimate the binding affinities of docking poses.

Methods
Our small-molecule docking algorithm FlexAID explores the conformational search space using an adaptive genetic-algorithm. It can use proteins, RNA or DNA as target and small-molecules or peptides as ligands. FlexAID is a coarse-grained docking algorithm as solvent molecules are treated implicitly and Hydrogen atoms are omitted. Thus, H-bond donors are assumed to be able to present a Hydrogen atom at any geometrically optimal position. The algorithm finds the optimal conformation of the complex by optimizing surface areas in contact between atoms in respect to their interaction2. Contact areas between atoms and solvent accessible surfaces are calculated using a Voronoi Polyhedron procedure3. The scoring function is a knowledge-based potential based on statistical analysis of pairwise atom interactions. We use sets of decoys selected from our unbiased dataset for training (57) and testing (60)4. A set of rules is used to assign one of the following types to non-hydrogen atoms: (I) Hydrophilic, (II) Acceptor, (III) Donor, (IV) Hydrophobic, (V) Aromatic, (VI) Neutral, (VII) Electrostatic, (VIII) Electrophilic, (IX) Positive or (X) Negative. The solvent is assigned type (I). The potential takes into account enthalpic (electrostatic interactions) and implicitly entropic effects as well. MM/GBSA re-scoring is used to estimate binding free energies. Additional degrees of freedom can be introduced to account for flexibility during simulations: 1) Flexible ligand bonds, 2) Side-chain flexibility using real rotamer instances5, 3) Coordinated global movements using our normal-mode analysis algorithm ENCoM. The performance of FlexAID is tested using a fixed number of energy evaluations against a subset of 157 proteins (bound and unbound forms) of our non-redundant dataset2.

Results & Conclusions
We compare the performance of FlexAID in self, cross and blind docking experiments. We show that most cases can be predicted accurately in the bound form. Our results also show that a fraction of cases could no longer be predicted when including side-chain flexibility. Moreover, we show that despite having no prior knowledge of the biological systems, we can still predict in a large fraction of cases the right conformation of the bound complexes.

We set up a grid computing infrastructure6 (NRG@Home), that allows us to dock thousands of compounds per hour and to perform high-throughput studies. We use a fragment-based drug design approach (FBDD)7, in the perspective of identifying lead compounds targeting the Germination Protease (GPR) of Clostridium difficile. We utilize the commercially available Enamine Golden Fragment Library containing 1190 fragments. Our results demonstrate the applicability of FlexAID to propose a pharmacophore model using FBDD (Figure 1) as the fragments with lowest estimated binding free energies show a core of conserved interactions.

References
THE QUIXOTIC QUEST FOR PERFECTION AT HIGH RESOLUTION

Daniel Keedy1, Michael Prisant1, Lindsay Deis1, David Richardson1*, & Jane Richardson1**
Dept. of Biochemistry, Duke University1. dcrfsr@kinemage.biochem.duke.edu

Experimentally determined macromolecular structures are often used as inputs to modeling applications such as drug design or protein engineering. However, all but a handful of the tens of thousands of deposited structures have locally misfit regions, even crystal structures at very high resolution. These errors may lead to faulty conclusions in prospective modeling efforts and hamper database-wide bioinformatics studies. Here we report on the identification and generation of “paragon” structures devoid of any demonstrably incorrect regions, which can serve as gold standards for modeling and motivate higher standards for macromolecular structure quality. We found proper treatment of alternate conformations to be a major bottleneck for producing paragons at ultra-high resolution, where many conformers are often visible in the density, and present ideas for automated modeling and labeling of self-consistent alternate conformation in the future.

INTRODUCTION
Deposited crystal structures contain more frequent problems than anyone would like to believe. Our statistics show that even at atomic resolution (1.4 Å or better), mean all-atom clashscore is 7.2, or one serious clash (probable fitting error) every 7 residues. 13% of Asn/Gln/His sidechains are clearly fit backward and 80% of files contain at least one case. These defects are one factor holding back the translational potential of molecular medicine.

Here we discuss “paragon” structures devoid of any such errors. Some already exist in the Protein Data Bank; many more can be created by remodeling a few regions in near-paragon structures. When collated, these structures can serve as a totally reliable test set for new modeling procedures.

In the process of creating paragons, we have also documented a number of complications involving alternate conformations at high resolution. Based on these observations, we also present some simple logical rules for defining self-consistent models.

METHODS
We pursued three approaches for obtaining paragon structures, all using our MySQL database containing MolProbity1 validation statistics. First, we queried for pre-existing paragons: deposited structures with no outliers whatsoever. Second, we queried for "cryptic" paragons: structures with only one or possibly two outliers that turned out to be strained but genuine conformations, as corroborated by local favorable interactions (e.g. hydrogen bonds), good electron density, and/or low B-factors. Third, we queried for structures with low numbers of outliers that turned out to be demonstrable fitting errors but seemed easily amenable to correction. Refittings were performed manually, followed by crystallographic re-refinement.

RESULTS & CONCLUSIONS
In the first category, we found only two ready-made paragons (first category): 2zqe and 3ifu. Given the fact that tens of thousands of crystal structures have been deposited to date, this result hammers home the difficulty of defining an entirely self-consistent and realistic model for a protein structure, even given the high information content of experimental data.

In the second category, Figure 1 shows a "cryptic" paragon: 3kyv. This case exemplifies the fact that outliers in our rotamer and Ramachandran distributions may (rarely) be valid conformations, if compensated by local stabilizing interactions.

In the third category, we found many candidates for “paragonization,” some easier to take to completion than others. On the easy extreme, 1akg required only a pucker flip of Pro 7 to better fit the Fo-Fc difference density and eliminate a steric clash. Somewhere in the middle of the difficulty scale, we made several changes to 1ubq to produce a paragon structure of ubiquitin, a 76-residue protein. On the difficult extreme, we made dozens of changes to 1gwe in an attempt to produce a paragon structure of catalase, which has 503 residues. These adjustments improved the MolProbity score1 from 1.42 to 1.00, although they failed to produce an entirely error-free structure.

REFERENCES
A NEW STRUCTURE AND SEQUENCE DEPENDENT NORMAL MODE ANALYSIS MODEL

Vincent Frappler, Jean-Guy Lehoux, Pierre Lavigne & Rafael Najmanovich

Dept. of Biochemistry, Sherbrooke University. Rafael.Najmanovich@USherbrooke.ca

Normal Mode Analysis is a computational technique used to generate an ensemble of movements for a given protein. Most known models are based only on the structure of the backbone, in opposition to our new model that also integrates the nature of the amino acid and their specific interactions. In comparison to two known models (STeM and ANM) we obtain better results in terms of structure prediction.

INTRODUCTION

Proteins are dynamic entities and instead of a single conformation (given by x-ray structure) they should be considered as an ensemble of conformations. Two main techniques are commonly used to evaluate the dynamic nature of proteins: nuclear magnetic resonance and molecular dynamic simulations. The first is a long and laborious experimental technique, while the second is a computationally demanding approach unable to produce large backbone conformational changes within the timeframe of standard simulations\(^1\). To overcome these problems, one alternative is to use Normal Mode Analysis (NMA). NMA explores the conformational space of a harmonic potential function by allowing the simultaneous movement of groups of atoms across the entire structure on a given resonance frequency\(^2\) (figure 1). Soft modes (slowest frequencies) are long timescale movements of low energy cost that may represent functionally relevant large movement. A common approach when predicting movement is to assume that the protein is already in an equilibrium conformation and to represent the whole amino acid using only the C\(_\alpha\) atoms. Despite the more simplistic representation, it was shown that such coarse-grained approaches give similar result as using all atoms\(^3\). The simplest coarse-grained model is the Anisotropic Network Model\(^4\) (ANM). It characterizes the interaction between the C\(_\alpha\) nodes within a predetermined cut-off distance with a simple Hooke type potential. On the other hand, the Spring Tensor Model\(^5\) (STeM) uses a more complex and realistic potential integrating the effect of covalent bond stretching, angle bending, dihedral rotation and Lennard-Jones interactions.

![Figure 1: Normal Mode Analysis on Barnase(1A2P.) Modes are presented as vectors applied on alpha carbon of the structure.](image)

METHODS

Our new model, called Elastic Network Contact Model (ENCoM), uses a four body potential (similar to STeM), but adjusts the long-range interactions by maximizing favourable contacts between interacting atoms (classified into 8 different atom types) based on surface complementary. Unlike other models, our approach takes in account side-chain orientation and the nature of amino acids. ENCoM could be able to differentiate, in terms of molecular movements, between two identical conformations (at the level of C\(_\alpha\) carbon atom positions) that differ in the identity of the actual amino acids present thus making it possible to study dynamic effects of mutations within normal mode analysis methods.

RESULTS & CONCLUSIONS

We tested ENCoM against ANM and STeM on the PSCDB\(^6\), a database of Apo-Holo protein pairs where ligand binding is associated to backbone conformational movements. We measured the overlap of the first 30 slowest modes to the vector describing the conformational change. The average best overlap shows that ENCoM (0.467) is better than ANM (0.433) and STeM (0.391) in terms of structure prediction. Moreover, we calculated the difference in vibrational energy between the wild type and the mutated form using the NMA predicted resonance frequencies. Over 17000 mutations were predicted from the ProTherm database. These mutations are stabilizing (37.9\%), destabilizing (35.9\%) or neutral (26.2\%) to protein stability. In terms of success rate, ENCoM (48.0\%, 1.4 fold over random predictions) is better in predicting the effect of mutations when compared to STeM (25.9\%) and ANM (25.2\%). As expected, STeM and ANM are not able to differentiate the effect of the mutations as these do not change the overall backbone structure and classify almost every mutation as neutral (over 92\%). Our results suggest a novel application for coarse-grained normal mode analysis methods in the prediction of the effect of mutations by considering the vibrational entropy change due to mutations.

In conclusion, the use of ENCoM allows us to study dynamic aspects of protein function without requiring long and laborious experimental or computational methods. Compared to other models it properly accounts for the chemical properties of amino acids. We are currently in the process of integrating ENCoM as part of our in-house developed genetic-algorithm based docking software, FlexAID, in order to introduce backbone movements into docking simulations.

REFERENCES

INTRODUCTION
Experimentally determined protein complexes are rare in comparison to the vast number of complexes that are found in systems biology experiments. Computational modeling and docking applications present a valuable alternative to conventional experiments. However, pure computational attempts often lack reliable and accurate predictions. On the other hand, data-driven modeling calculations promise high-quality models, but rely on correct experimental data, of which there are none for most proteins. Here, we present a hybrid structural biology approach that is based on chemical cross-linking experiments and state of the art molecular modeling and cross-link visualization software, which was applied to gain structural insights into the protein interaction network of human PP2A. PP2A is a trimeric serine/threonine phosphatase, consisting of a scaffold, a regulatory and a catalytic subunit, which is involved in various cellular functions—among others, cell-growth control and regulation of signal transduction cascades.

METHODS
Chemical cross-link data between lysine residues were generated on purified human PP2A complexes, using isotopically labelled disuccinimidyl suberate (DSS) as a cross-linking reagent and liquid chromatography coupled to a tandem mass spectrometer. Cross-linked peptides were identified with the search engine xQuest. Based on the length of DSS a maximum threshold of 30.0 Å was defined for Euclidean distances between the cross-linked lysine Cα atoms. In addition, Solvent Accessible Surface Distances (SASD) with a maximum threshold of 34.0 Å were calculated using Xwalk and utilized to remove false positive Euclidean distance measurements.

Comparative modeling, de novo modeling, loop modeling and protein-protein docking were performed using the ROSETTA modeling suite. Cross-link data were supplied as Euclidean distance constraints to the ROSETTA scoring function using a flat/harmonic function and employed as SASD filter on predicted models. Models passing the filter were clustered and decays with the lowest ROSETTA scores from the largest clusters were selected as best models.

RESULTS & CONCLUSIONS
Cross-links were validated on existing crystal structures of trimeric PP2A complexes (3fga.pdb, 3dw8.pdb) and comparative models of isoforms. Over 90% of cross-links were found to have distances below the maximum distance threshold of 30.0 Å.

A full length model of human IGBP1, which forms a binary complex with the catalytic subunit of PP2A, was constructed based on 65 intra-protein cross-links using comparative modeling on the N-terminus and de novo modeling on the C-terminus (see Figure 1). The full length model conforming to 60 out of 65 cross-links was subsequently docked with the catalytic subunit of PP2A employing 7 inter-protein cross-links (see Figure 1). The most frequent interface residues within the ensemble of best models agreed with residues described as important for the interaction of IGBP1 and the catalytic subunit of PP2A.

Furthermore, the positively charged N-terminal region of SGOL1 was modeled by loop-modeling using 11 intra-protein cross-links and 2 inter-protein cross-links to the catalytic and the regulatory subunit of PP2A. One of the best models showed a helix-helix interaction between the N-terminal SGOL1 helix and the negatively charged HEAT repeats of the regulatory subunit (see Figure 1).

The ability to produce distance information in large quantities on systems level and to employ these in modeling calculations is a promising technology for gaining structural insights into interaction networks that are difficult to elucidate otherwise.

Figure 1. PP2A complex (transparent) with scaffold (grey), regulatory (orange) and catalytic subunit (cyan) as described in 3fga.pdb. Full length model of IGBP1 (magenta) conforming to 60 intra-protein cross-links was docked to the catalytic subunit using 7 inter-protein cross-links. Crystallized SGOL1 peptide (yellow) was extended towards the N-terminus (blue) using 2 inter-protein cross-links and 11 intra-protein cross-links.

REFERENCES
**THE DEPLETION OF CALCIUM IONS EXPLORED BY MD SIMULATIONS – MULASE STUDY ON THROMBOSPONDIN**

Thomas Haschka1,2,*, Catherine Etchebest1, Eric Henrot3, Laurent Martiny1 & Manuel Dauchez2
1SIRMA Laboratory CNRS UMR6237 - Matrice Extracellulaire et Dynamique Cellulaire (MEDyC)
2ICMR UMR CNRS 7312
3Université de Reims Champagne-Ardenne, Pôle Sciences, Moulin de la Housse - BP1031, 51687 Reims Cedex 2

Calcium ion-protein interactions play an important role in several biological processes. Several calcium binding proteins do exist and many of them share the DxDxDG motive. One of them is thrombospondin with its calcium binding signature domain. Simulating the process of calcium depletion/repletion is a complicated task to achieve with today’s non-polarizable force fields. Nevertheless we tried to approach this problem by performing simulations of the thrombospondin signature domain under various force field parameters, revealing interesting insights on calcium depletion and the inherent conformational change caused by it.

**INTRODUCTION**

Calcium ions are involved in many biological processes and interact with diverse proteins in specific binding sites. EF hands are well-known examples of protein domains in where the presence or absence of calcium can induce major conformational changes.

The thrombospondin (TSP) family is another important example of Calcium binding proteins. TSPs contain different Ca-Binding domains, EGF-like and a signature domain which comprises N and C-like repeats containing the DXDx(D)G motive. They are implicated in various biological functions such as wound healing, inflammation, angiogeneses, cartilage formation, proliferation, chondrogenesis and neoplasia. All five members of the TSP family contain a highly conserved signature domain (SD), consisting of several calcium-binding sites. The SD was found to undergo conformational changes depending on the calcium concentration of the solute. All crystal structures that have so far been resolved are only available in calcium replete conformations. As the calcium depleted form is supposed to be functionally relevant we tried to tackle calcium depletion by the use of molecular dynamics simulations. However, validity of classical force fields for describing interactions of macromolecule with calcium ions and divalent ions in general, remains subject of debate.

In the present work, we have explored different TSP constructs, which are replete in Ca ions. Using MD simulations and testing different force fields, we were further able to follow the ions motions and how they impact on the protein’s conformation.

**METHODS**

The GROMACS software package was used to perform 48 MD simulations of 50ns of various structures from the thrombospondin signature domain. Simulations were performed either using i) the GROMOS 53a6 force field (g53a6), or an improved version of Asp-Ca2+ interactions and ii) the OPLS all-atom force field.

Going further we investigated the calcium binding sites at the quantum mechanics level using GAUSSIAN software, to gain insights on how charges are modified on the protein in areas of calcium binding.

**RESULTS & CONCLUSIONS**

We showed that ions behave differently depending on the nature and the location of the binding sites in the structure. We identified the location of SD binding ions that are probable to detach themselves from the protein. Further we gained insights on the calcium depletion process itself and were able to show how local binding sites are effected by the loss of their calcium ions. We also demonstrated the influence of ions on protein flexibility.

Hence, we gained several insights on the process of calcium depletion itself. This study is even more important as many calcium binding proteins feature the significant DxDxDG motive found in the SD of TSPs.

As a byproduct we gained insights on the flexibility of the signature domain and propose new models for the calcium depleted form of the protein.

**REFERENCES**

IDENTIFY DRUGGABLE TARGETS USING MICROENVIRONMENTS

INTRODUCTION
The ability of a target protein to bind small, drug-like molecules with a high affinity is referred to as druggability. It is an important criterion in the target selection phase for drug discovery. Despite numerous screening attempts, many targets have failed to show any evidence of binding drug-like molecules. Known small molecule drugs occupy a limited area of chemical space and therefore their binding sites should share common features. Therefore through structural analysis, it is possible to filter binding sites with their inherent ability to bind high affinity, drug-like molecules. However, there are only a few studies addressing methods for evaluating target druggability. The first experimental assessment of protein druggability, proposed by the Abbott laboratories, relies on the 2D NMR screen of a fragment library in which hit rates were demonstrated to correlate with the protein ability to bind drug-like ligands with high affinity, and thus proposed as a reliable indicator of protein druggability. The theoretical estimation of druggability using a model-based approach was performed by Cheng and co-workers. They quantified the maximal affinity achievable by drug-like molecules for druggable and undruggable targets and found that the calculated affinity correlates with drug discovery outcomes. As an alternative to executing an NMR-based screening against drug fragment libraries, we developed a novel computational method, named DrugFEATURE, to calculate target druggability and predict candidate drug/drug fragments for repurposing.

METHODS
Given a functional center of a residue, we use the term “microenvironment” to refer to the local, spherical region in the protein structure that may encompass residues discontinuous in sequence and structure. Specifically, we use the FEATURE system to calculate a set of 480 physicochemical properties collected over the predefined functional center. We previously developed a scoring system to calculate similarities between microenvironments from two sites and identify microenvironments that confers molecular recognitions. In this work, we created a dataset of the 3D structures of known drug-binding sites, representing the optimal microenvironments for drug binding. Given a target, we compare it to the representative of drug-binding sites and identify similar subsites (clusters of key microenvironments). These similar subsites confer molecular recognition of chemical groups of drugs. DrugFEATURE calculates the frequency of similar subsites and derives druggability of the given target (Figure 1).

RESULTS & CONCLUSIONS
We benchmarked DrugFEATURE using the Abbott dataset of which the hit rates of NMR screening were made available. The NMR hit rates have been demonstrated to correlate with the ability to bind drug-like molecules with high affinity. In our calculation, the DrugFEATURE druggability shows correlation with NMR-based hit rate (Figure 1). DrugFEATURE druggability also effectively discriminates druggable sites from undruggable ones.

REFERENCES
A RATIONAL PROTEIN REDESIGN METHOD FOR IMPROVED SECRETION YIELDS

B.A. van den Berg1,2, M.J.T. Reinders1,2, J.M. van den Laan1, J.A. Roubos1 & D. de Ridder1,2,4
1Delft Bioinformatics Lab, Delft University of Technology, 2Netherlands Bioinformatics Centre, 3DSM Biotechnology Center, Delft, and 4Kluiver Center for Genomics of Industrial Fermentation,
b.a.vandenberge@tudelft.nl

A rational protein redesign method is developed to alter sequence properties without affecting enzymatic function. The rationale behind this is the observed characteristic amino acid composition for proteins that provide high extracellular concentrations after over-expression compared to those that do not; thereby suggesting that alteration of this characteristic can be used to improve secretion yields. However, the enzymatic activity should remain similar to the original activity to make the method useful for the enzyme industry. Here, as an in silico validation, it is shown that our method is able to increase the sequence similarity of a protein with a low secretion yield to a close homolog with a high secretion yield. Experimental validation is being performed to validate our method and hypothesis in vivo.

INTRODUCTION

In industrial biotechnology, enzymes are produced by over-expressing genes in production hosts such as Aspergillus niger. Ideally, the produced protein should be secreted, such that it can easily be recovered from a reactor. Previously, we exploited a dataset of proteins to predict successful high-yield secretion (1), showing that the frequency of occurrence of the different amino acids is most predictive. Here, we develop a method to rationally redesign the sequence of a low-yield protein to make the amino acid composition more similar to that of high-yield proteins (Figure 1), with the aim to optimize the secretion. Our method was validated in silico, by showing that the redesigned sequence has an increased sequence similarity to a structurally similar high-yield protein that was not used in the design process. Experiments are currently being performed to also validate the method in vivo.

METHODS

The method optimizes a fitness score that combines three objectives. First, mutations from amino acids with a low weight to amino acids with a high weight are promoted, in which the amino acid weights are obtained from the trained predictor. This increases the protein's correspondence to high-yield proteins. Second, mutations to amino acids that are frequently observed at the same position in homologous proteins are preferred, thereby providing basic confidence that the mutation will not corrupt protein structure and activity. Finally, mutations are optimized for an amino acid distribution that is close to the average distribution of the high-yield proteins. As an additional constraint, residues in the vicinity (< 8Å) of the active site and residues for which the relative accessible surface area is less than 5% are fixed, i.e. not allowed to be mutated. A genetic algorithm optimizes the combined fitness score for a fixed number of mutations.

RESULTS & CONCLUSIONS

The design method is applied to a protein (wt) for which a low secretion yield was observed and that has a close homolog for which a high yield was observed (wtµ). This allows us to test whether the sequence of the redesigned sequence (wtred) becomes more similar to that of wtµ.

For a design run in which 30 residues were mutated, 16 identical residues in the structural alignment with wtµ were gained, while 2 were lost (blue dots in Figure 2), i.e. about half of the mutations is to an amino acid that is identical to the one found at the same position in wtµ. Side chains of two mutated residues are shown as example in Figure 3. In both cases the mutated side chain (green) is positioned similar to the one in the aligned structure of wtµ (blue). We are currently testing secretion of a number of proteins redesigned using our method in the lab.
PREDICTION OF REAL-VALUE FLUCTUATION OF GLOBULAR PROTEINS

Michal Jamroz1,2, Andrzej Kolinski1, Daisuke Kihara3,4,5
Faculty of Chemistry, Warsaw University1, Depts. of Biological Sciences2 and Computer Science3, Purdue University. dikihara@purdue.edu

Flexibility and dynamics is very important for protein functions and interactions. Here we developed a computational method that predict real-value of fluctuation at each amino acid of a protein with an average Pearson’s correlation coefficient of 0.669 and a root mean square error of 1.04 Å.

INTRODUCTION
It is crucial to consider dynamics for understanding biological function of proteins. We used a large number of molecular dynamics trajectories of non-homologous proteins for investigating static structural features of proteins that are most relevant to fluctuations. We examined correlation of individual structural features with fluctuations and further investigated effective combinations of features for predicting the real-value of residue fluctuations using the support vector regression (SVR). It was found that some structural features have higher correlation than crystallographic B-factors with fluctuations. Moreover, SVR that uses combinations of static structural features showed accurate prediction of fluctuations with an average Pearson’s correlation coefficient of 0.669 and a root mean square error (RMSE) of 1.04 Å. This correlation coefficient is higher than the one observed for the prediction by the Gaussian network model (GNM). An advantage of the developed method over the Gaussian network models is that the former predicts the real-value of fluctuation. The results help improve our understanding of relationships between protein structure and fluctuation. Furthermore, the developed method provides a convenient and practical way to predict fluctuations of proteins using easily computed static structural features of proteins.

METHODS
In total of 837 molecular dynamics (MD) trajectories of nonhomologous proteins were selected from the MoDEL database2 for this study. The trajectories were computed using AMBER, GROMACS or NAMD MD suites. The simulation time was 10 ns for most of the proteins (96.11%), while the rest of the proteins had shorter trajectories. The fluctuation of each residue during MD simulation was computed to relative to the position in the crystal structure in the PDB file of the proteins.

We examined the following eight types of structural features of residues in proteins in terms of correlation and predictive performance of residue fluctuation: 1) the crystallographic B-factor; 2) the distance to the center of mass of the protein; 3) the number of contacting residues using eight cutoff values of Cα-Cα distance; 4) the number of residue contacts in lower/upper half-sphere of residues; 5) the accessible surface area; 6) the residue depth; 7) secondary structure; as well as 8) fluctuations predicted by the GNM. Except for the last feature, all the others are static feature of protein structures.

RESULTS & CONCLUSIONS
First, we compared the correlation coefficient of individual structural features with the fluctuation of residues in the MD trajectories. Interestingly, several static structural features, namely, the distance to the center of mass (0.51), the contact number computed with the cutoff of 12-22 Å (-0.55 to -0.57), have more significant correlation than the B-factor (0.48). The correlation coefficient values are shown in the parentheses. Among the static features, the largest correlation coefficients were observed for the residue contact number (15 and 16 Å (-0.57). These results indicate that the motion of protein chains is better captured by the coarse-grained topological structures of proteins rather than the B-factor.

Next, we employed SVR to predict the residue fluctuation in the MD trajectories using various combinations of static structural features (without incorporating GNM). Five-fold cross validations were performed. Among the seventeen feature combinations tested, all except for two showed a higher correlation with actual fluctuations than GNM showed. The largest correlation coefficient, 0.669, was achieved when the residue contact numbers with different distance cutoffs were combined as features. In terms of average RMSE, all the feature combinations predicted residue fluctuations within an RMS of 1.1 Å, ranging from 1.042 Å to 1.092 Å.

The current study demonstrates that flexibility of proteins is inherently coded in coarse-grained static protein structural features, even more than in the crystallographic B-factors. Thus, protein motion is determined by its static structure that is coded by its sequence, which could be considered as an extension of the Anfinsen’s dogma. Indeed series of studies on GNM has also demonstrated that motion of a protein is determined by its structure. The current work further shows that static structural features can predict the real-value of fluctuations, which GNM has not been shown to be able to do.

Funding: This work is supported partly by grants from NIH (R01GM075004, R01GM097528) and NSF (IOS1127027, EF0850009, IIS0915801).

REFERENCES
In the field of molecular modeling, normal modes analysis (NMA) has been shown to be an effective computational method to study the movements of proteins, especially at the domain level. The lowest frequency modes have been found to correlate well to functionally relevant protein motions. Moreover, coarse-grained NMA has been shown to agree well with computational expensive all-atom NMA and molecular dynamics simulations. We and others have that the scope of using NMA to analyse structures is immense in terms of looking at protein dynamics. Since 2000, several NMA web-servers have been created but none compare the normal modes of a protein with different aligned structures. For simple and easy access to this method, we improved WEBnm@ (http://apps.cbu.uib.no/webnma/home), a web-tool that provides access to NMA on coarse-grained models of protein structures and various analyses. The output is available and downloadable as both R² plots or raw data points. The animations of some of the lowest modes are visualised interactively using Jmol.

**METHODS**

The normal modes analysis calculations are done using Molecular Modeling Toolkit (MMTK)², where the Cα forcefield is used. In the Single Analysis section we have added, an interactive visualisation (through a Jmol applet) of the six lowest frequency modes (both vectors and animation), the calculation of the correlation matrix based on all the modes and the overlap analysis with another conformation of the same structure. The newest section, Comparative Analysis, calculates and compares the normal modes of a set of aligned protein structures. For this part, more than one structure can be submitted along with a FASTA file of their alignment. It includes analyses such as the fluctuations profile and deformation energies as they are calculated in the Single Analysis section but produced in a way that reflects their sequence alignment. The conformational overlap analysis is done by calculating the root mean squared inner product (RMSIP) of the lowest ten modes of the fully conserved parts of the sequence alignment of all the structures submitted. In addition to the updates, we provided a Simple Object Access Protocol (SOAP) web-service using for programmatic access to both of these sections.

**RESULTS & CONCLUSIONS**

To demonstrate the newly added functionalities, we submitted Sarcoplasmic/endoplasmic reticulum calcium ATPase (SERCA1) as a case study to both the Single and Comparative Analysis sections. SERCA is known to undergo large conformational changes to transport calcium ions across the cell membrane and has been previously studied using NMA. Through this example, we show that WEBnm@ is useful for exploratory NMA and allows for easy comparison of more than one structure. It provides analyses not seen in other web-tools, in particular, the Comparative Analysis. We have improved the efficiency for input processing and have added a clearer error/warning system at the users’ end. The data files produced are available for download and there is a programmable web-service that allows users to manipulate the options available on WEBnm@ to suit their needs.

**REFERENCES**


**FIGURE 1.** Front end of WEBnm@. The yellow box shows the tabs for the Single and Comparative analysis input fields.
ON THE FLY – A DATABASE OF DROSOPHILA MELANOGASTER TRANSCRIPTION FACTORS (TF) AND TF BINDING SITES (TFBS)

Shula Sharman1,2, Jie Chen2,3, Hanjoong Lee2,3, Richard Mann1 & Barry Honig1,2,3

1Department of Biochemistry and Molecular Biophysics, Columbia University, New York, 2Center for Computational Biology and Bioinformatics, Columbia University, New York, 3Howard Hughes Medical Institute; bh6@columbia.edu

We describe a systematic determination of Drosophila melanogaster transcription factor DNA-binding specificities. We annotated and classified all Transcription Factors (TFs) predicted in the genome and collected the known preferred DNA binding sites of the TFs based on the B1H, DNaseI and SELEX experimental methods. Then, we identified the sequence and shape preferences for all DNA binding proteins and also characterized the shapes of their preferred DNA binding sites using structural models. The identification of the preferred DNA binding sites, and their shapes, for all DNA binding proteins will provide an unprecedented and extremely valuable database for anyone attempting to decipher non-coding regulatory DNA. Furthermore this is the first time that all DNA binding domains encoded in a eukaryote are fully annotated using structural criteria and evolutionary homology.

INTRODUCTION

Despite their abundance and central importance to biology, we are still largely ignorant about how regulatory DNAs are interpreted in living cells. The interpretation of regulatory DNAs depends on the binding of proteins to these sequences. The goal of this study is to assemble sequence and structural information, at an unprecedented level of resolution, for all DNA binding proteins encoded in the Drosophila melanogaster genome. We chose D. melanogaster because its genome encodes multiple members of all known families of DNA binding proteins, yet the total number of TFs is manageable (~754). There are many databases which house collections of TF DNA binding information. However, there is currently no organism for which a complete encyclopedia of such TF sequence and structure specificities exists. Recent studies revealed a previously under-appreciated role for DNA shape, or DNA topography, in protein-DNA recognition [1]. With this in mind, we characterized the DNA topographies of the binding sites identified by the B1H, DNaseI and SELEX methods.

METHODS

Dataset collection

The list of “all predicted TFs” listed on the Flytfl[2] web site was chosen as the basis for our dataset, since we wanted to be as comprehensive as possible. This list contains 754 genes which produce 2107 proteins. PWMs were obtained from Casey Bergman’s FlyReg dataset of DNase I footprint experiments [3], and from FlyFactorSurvey web site based on B1H and SELEX methods [4].

DNA model

To study sequence-dependent DNA shape, we developed a course grained model (to be published). The corresponding force field is derived from all-atom molecular dynamics simulations trajectories previously obtained.

Protein model

The basic idea of study sequence-dependent protein shape is to produce constraints for dynamic programming using structural alignments of a template structure and a homology model. The HHpred (or HHMER) is used for template selection, MODELLER for building models, PROSA-II for model evaluation. Kuziemko et al recently have developed the algorithm to explore the sequence alignment space, S4 (Sampling Shifts in Secondary Structures) [5]. S4 produce constraints for dynamic programming.

RESULTS & CONCLUSIONS

The database provides, for each TF, the functional domain, 3D structural model and the known DNA sequence this TF prefers to bind (Figure 1) based on experimental methods.

FIGURE 1. Three families of TFs and their DNA binding specificities.

For each DNA sequence we extract and present DNA structural features using the Curves program. Using the DNA sequences stored in the database we show that using structural criteria such as the width of minor groove we are able to distinguish DNA sequences bound by homeodomains from sequences that bind to ZNF-C2H2 domains or ETS domains, with 80% accuracy. This separation cannot be achieved using sequence motifs alone. Our results make it clear that an understanding of protein-DNA recognition cannot simply treat DNA sequences as linear strings of nucleotides, but that the 3-dimensional structures of the DNA molecules must also be taken into consideration.

REFERENCES

MEGADOCK: A RAPID SCREENING SYSTEM OF PROTEIN-PROTEIN INTERACTIONS BY EXHAUSTIVE DOCKING

Masahito Ohue1, Yuri Matsuzaki1, Nobuyuki Uchikoga1, Takashi Ishida1 & Yutaka Akiyama1,2
1Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2Research Fellow of the Japan Society for the Promotion of Science
3Dept. of Physics, Faculty of Science and Engineering, Chuo University
akiyama@cs.titech.ac.jp

We developed high-speed all-to-all protein-protein interaction prediction system, called MEGADOCK. We applied MEGADOCK to the general benchmark dataset and the bacterial chemotaxis signaling pathway.

INTRODUCTION

Protein-protein interactions (PPIs) occur when two or more proteins bind together, often to carry out their biological functions. PPIs have been extensively investigated from the perspectives of biochemistry, quantum chemistry and molecular dynamics. PPIs are at the core of cellular processes of any living cell, and investigating PPIs is crucial for understanding cellular processes. As a result, several methods to determine protein pairs that have potential to bind and interact with each other; this is called PPI screening.

In this study, we have developed a rigid-body docking-based method of PPI screening based on exhaustive calculations of pseudo-binding energies among pairs of target proteins. Further, in order to enable applications to PPI prediction problems of mega-order data, we have developed a fast protein-protein docking software aimed at exhaustive PPI screening. Our PPI screening system is called MEGADOCK.

RESULTS

Figures 1 shows the overall procedures for PPI prediction by employing the MEGADOCK system. MEGADOCK consists of two parts—a “docking calculation” part and “PPI decision” part. The “docking calculation” part performs all-to-all docking and generates high-scoring decoys for all possible combinations of the given protein structures. Subsequently, the “PPI decision” part analyzes the structural distributions of high-scoring decoys for each pair of proteins and decides if the given two proteins can interact. We have been investigating a protein-docking approach based on shape complementarity and physico-chemical properties. To realize the procedures required to sample a huge number of protein dockings, we have developed a novel scoring function called the real-Pairwise Shape Complementarity (rPSC) score. We demonstrate that MEGADOCK is capable of exhaustive PPI screening by completing docking calculations 7.5 times faster than the conventional docking software, ZDOCK, while maintaining an acceptable level of accuracy.

When our PPI prediction system was applied to a subset of a general benchmark dataset to predict 120 relevant interacting pairs from 120 x 120 = 14,400 combinations of proteins, an F-measure value of 0.231 (recall 0.150, precision 0.500) was obtained.

In addition, as a case study using real biological data, we also applied our method to the protein structures of the bacterial chemotaxis pathways, which represents a typical target in the systems biology field. Although proteins included in these pathways are not representing all the known proteins in biology, the pathways have been extensively studied over the past several decades and thus we assumed that most of the essential protein-protein interactions were known. So we used this system as a target to evaluate our system on a real biological problem. We used 89 protein structures which corresponded to 13 protein species (CheA, CheB, CheC, CheD, CheR, CheW, CheX, CheY, CheZ, Tsr, FliM, FliG, FliN) and obtained an F-measure score of 0.364 (recall 0.300 precision 0.462) for this system.

CONCLUSIONS

MEGADOCK showed comparable docking accuracy to other FFT-based software programs, such as ZDOCK, while employing a much simpler and thus computationally less expensive score function. Additionally, our software was shown to be applicable to a large scale protein-protein interaction screening problem with accuracy better than random. With our approach combined with parallel high-performance computing systems, searching and analyzing protein-protein interactions with consideration to three-dimensional structures at the interactome scale is now a feasible problem.

REFERENCES

We identify interactions between membrane transporters and small molecule ligands, including prescription drugs and metabolites, using comparative modeling, virtual screening, and experiment. Our results may explain some of the pharmacological effects (ie, efficacy and/or side effects) of these drugs via polypharmacology. For example, the large neutral amino acid transporter (LAT-1) is a transporter found in the Blood-Brain Barrier (BBB), where it transports metabolites and drugs into the brain. We modeled LAT-1 based on the structure of the amino acid antiporter AdiC, and performed virtual screening of ~20,000 drugs and metabolites; five of the high scoring molecules were confirmed experimentally as LAT-1 ligands, including two chemically novel ligands and two substrates, which also get transported across the membrane via LAT-1 (Figure 1). These results may rationalize the enhanced brain permeability of the antiepileptic drug Rufinamide.

Second, applying our combined modeling/docking approach to membrane proteins that share less than 30% sequence identity with their template structures, we correctly predicted chemically novel SLC ligands. This suggests that our structure-based approach might be useful for identifying unknown interactions between proteins and novel small-molecule ligands.

**RESULTS & CONCLUSIONS**

Three key findings emerge from our studies. First, several drugs and metabolites that target a variety of receptors and enzymes are also ligands of SLC transporters. This might explain the pharmacological effect of these molecules (ie, efficacy and/or side effects), and also demonstrates the prevalence of polypharmacology. For example, the large neutral amino acid transporter (LAT-1) is a transporter found in the Blood-Brain Barrier (BBB), where it transports metabolites and drugs into the brain. We modeled LAT-1 based on the structure of the amino acid antiporter AdiC, and performed virtual screening of ~20,000 drugs and metabolites; five of the high scoring molecules were confirmed experimentally as LAT-1 ligands, including two chemically novel ligands and two substrates, which also get transported across the membrane via LAT-1 (Figure 1). These results may rationalize the enhanced brain permeability of the antiepileptic drug Rufinamide.

Second, applying our combined modeling/docking approach to membrane proteins that share less than 30% sequence identity with their template structures, we correctly predicted chemically novel SLC ligands. This suggests that our structure-based approach might be useful for identifying unknown interactions between proteins and novel small-molecule ligands.

**REFERENCES**

A PROBABILISTIC FRAGMENT-BASED PROTEIN STRUCTURE PREDICTION ALGORITHM

David Simoncini1, Francois Berenger1, Rojan Shrestha2& & Kam Y. J. Zhang1,2,*
1Zhang Initiative Research Unit, Advanced Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan; 2Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan. kamjang@riken.jp

A method based on Estimation of Distribution Algorithm is developed for protein structure prediction using fragment assembly approaches. When compared with Rosetta AbInitio protocol on a benchmark of 20 proteins, the result shows that our method is able to generate models with lower energies and to enhance the percentage of near-native coarse-grained decoys on these proteins. This has been shown to lead to improved all-atom models that could be used as templates to solve the crystallographic phase problem by molecular replacement.

INTRODUCTION
Conformational sampling is one of the bottlenecks in fragment-based protein structure prediction approaches. They generally start with a coarse-grained optimization where mainchain atoms and centroids of side chains are considered, followed by a fine-grained optimization with an all-atom representation of proteins. It is during this coarse-grained phase that fragment-based methods sample intensely the conformational space. If the native-like region is sampled more, the accuracy of the final all-atom predictions may be improved accordingly.

Amongst all the de novo structure prediction methods developed, Rosetta is one of the most well-known [1]. The Rosetta protocol consists of a coarse-grained optimization phase where backbone atoms and centroids of side chains are considered, followed by a fine-grained optimization phase with a high resolution all-atom representation of proteins. The fragment assembly takes place during the first phase. Short fragments of known proteins are assembled by a Monte Carlo strategy to generate decoys.

The two main challenges in protein structure prediction are on the one hand the accuracy of energy functions and on the other hand the efficiency of the conformational search. The conformational search space is determined by the energy function and the representation of the solutions. Despite of its imperfections, Rosetta’s all-atom energy function can distinguish native from non-native structures [2]. The challenge we are interested in is, given an accurate energy function, to increase the proportion of near native decoys during the conformational search. To achieve this goal, an optimization method must have features that can bring some knowledge on the search space. Resampling techniques and methods inspired by evolution such as genetic algorithms possess this kind of features and thus are good candidates for this task.

In this work we present EdaFold, a new method for fragment-based protein structure prediction based on an estimation of distribution algorithm. Fragment-based approaches build protein models by assembling short fragments from known protein structures. Whereas the probability mass functions over the fragment libraries are uniform in the usual case, we propose an algorithm that learns from previously generated decoys and steers the search towards near-native regions.

METHODS
The corner stone of the proposed algorithm is the estimated probability mass functions (PMFs) defined over the protein fragment library. We define a probability mass function for each residue in the target sequence. At the beginning, no information on the fitness of fragments to the sequence is available and thus every fragment has the same probability of being inserted into a solution. The PMF estimations are then refined after each sample and minimization process.

The estimation of PMFs requires to be able to trace the mainchain atoms and centroids of side chains for every possible fragment. It is during this coarse-grained phase that fragment-based methods sample intensely the conformational space. If the native-like region is sampled more, the accuracy of the final all-atom predictions may be improved accordingly.

However, the negative impact of such landscapes on performance is limited. The process we have implemented mitigates the loss of diversity in the final population and avoids from drifting too far from native structure in case of deceptive landscapes. For some protein targets, EdaFold can achieve better results than Rosetta even when the energy landscape is deceptive.

REFERENCES
ORGANIZING MOLECULAR PRINCIPLES OF THE MOLECULAR CHAPERONE DYNAMICS AND REGULATORY INTERACTIONS

Gennady Verkhivker1,2*, Nathan Lawless3, Elizabeth Berrigton, Kristin Blacklock1
Dept. of Computational Sciences, Chapman University4 and Dept. of Pharmacology, University of California San Diego5. verkhivk@chapman.edu

By integrating structural bioinformatics, biophysical simulations, modelling of allosteric communications and the energy landscape theory we characterize functional dynamics and regulatory interactions of the Hsp90 molecular chaperones. This approach has identified a network of conserved regions that could play a universal role in coordinating functional dynamics, allosteric signalling and protein client binding of Hsp90.

INTRODUCTION

The molecular chaperone Hsp90 (90 kDa heat-shock protein) is an important hub in a variety of protein interaction networks associated with oncogenic pathways and responsible for the conformational maturation of proteins. Structural and biochemical studies of Hsp90 have proposed that the molecular chaperone could function via a nucleotide-triggered switching between the apo-open and ATP-closed states, wherein ligand-based modulation of the conformational dynamics can bias the equilibrium towards functionally relevant complexes. According to this mechanism, the ATPase cycle of Hsp90 is not a conformationally deterministic process, but rather a functional regulator of the equilibrium between pre-existing conformational states.

METHODS

MD simulations were carried out using NAMD 2.6. The employed protocol was described in full details in our earlier studies. Protein dynamics was analysed using the principal component analysis and energy landscape models. To evaluate communications between protein residues, we computed the communication propensity between all protein residues that can be defined as the mean-square fluctuation of the inter-residue distance. Exploiting a spectrum of reference threshold distances (20Å-70 Å), we have then computed the “long-range communication capability” values for each protein residue, defined as a fraction of residues that efficiently communicate at distances larger than the threshold. The resulting “scanning histograms evaluate communication efficiencies given by the fraction of residues that have high communication efficiency at distances larger than the cut-off.

RESULTS & CONCLUSIONS

Despite significant advances, organizing molecular principles that control the relationship between conformational diversity and functional dynamics of Hsp90 lack a sufficient quantitative characterization. We have reported a series of computational studies of the Hsp90 allosteric binding with an atomic level analysis of the inter-domain communication pathways which may regulate the conformational equilibrium of the molecular chaperone. We have found that in the presence of ATP long-range communication from the nucleotide binding site is mainly directed to the specific residues at the CTD dimerization interface, while ADP can activate allosteric signalling between the binding site and residues from the C-terminal region that surround the CTD dimerization interface. These results have provided evidence of a possible cross-talk between N- and C-terminal binding sites of Hsp90 that may induce an allosteric regulation of the molecular chaperone machinery. By integrating structural bioinformatics, biophysical simulations, modelling of allosteric communications and the energy landscape models and structure-functional analyses, we characterize functional dynamics and regulatory interactions of the Hsp90 molecular chaperones. We identified and characterized evolutionary and functionally conserved elements of the Hsp90 chaperone that may serve as key regulators of collective motions and hubs of long-range communication networks. We have found that functional sites of Hsp90 involved in allosteric signaling, catalysis and binding may be strategically positioned near key anchoring regions and inter-domain interfaces to control global movements of the molecular chaperone. The central result is that a small number of functional motifs may be utilized by the chaperone machinery to act collectively as central regulators of Hsp90 dynamics and activity, including the inter-domain communications, control of ATP hydrolysis, co-chaperone binding and protein client recognition.

REFERENCES

FITTING MULTIMERIC PROTEIN COMPLEXES INTO ELECTRON MICROSCOPY MAPS

Juan Esquivel-Rodríguez & Daisuke Kihara1,2

INTRODUCTION

Highly complex functions in the cell depend on large macromolecular machines, like GroEL/GroES, ribosomes and nuclear pore complexes. Electron microscopy (EM) has been proven very useful to determine the overall structure of these large assemblies. However, to analyze the functional mechanisms of macromolecular complexes it is desirable to integrate atomic-detailed structures along with low-resolution EM maps.

We have developed EMLZerD1 that fits multiple high-resolution structures into an EM map. It combines a multiple protein docking procedure (Multi-LZerD)2,3 and surface shape comparison between protein complex structures and the EM map using the 3D Zernike descriptor (3DZD). The 3DZD provides a unified representation of the surface shape of multimeric protein complex models and EM maps, which allows a convenient, fast quantitative comparison of the 3D structural data. Existing methods focus on local optimization and require placing component proteins at pre-assigned anchoring positions in the EM map. In contrast, our approach does not need initial pre-assignment of proteins and explore a larger complex conformational space. Moreover, the shape comparison using the 3DZD can tell if correct docking solutions exist among the candidates constructed or not.

RESULTS & CONCLUSIONS

The 3DZD provides a unified representation of the surface shape of multimeric protein complex models and EM maps that allows a convenient, fast quantitative comparison. In the benchmark of 19 multimeric complexes, near-native structures were obtained at both EM map resolutions studied. EMLZerD was able to identify near-native complex structures (2.5 Å or closer to native) for 14 cases and medium range models for the others. Among the test cases, there are few cases where the method did not yield a near-native model. However, the 3DZD can indicate if a sufficiently accurate model is included in the pool of generated candidate structures or not.

As more protein complexes are being solved by EM, it is crucial to provide computational methods that aid in the analysis of low-resolution structures from EM. EMLZerD, together with the other existing methods, will be a valuable tool for EM structural biology.

Funding: This work is supported partly by grants from NIH (RO1GM075004, ROI1GM097528) and NSF (IOS1127027, EF0850009, IOS1195801).

REFERENCES
INTRODUCTION
As the number of protein structures in the Protein Data Bank is rapidly growing, new ways to identify structural and functional similarity are needed. It is well established that the surface of proteins plays a significant role in defining their function. Recent studies have classified protein surface patches, based on their geometric and/or electrostatic features [1-3]. Here, we present PatchBag – a novel approach to characterize the entire protein surface by the distribution of all its fixed-size overlapping surface patches. We show that PatchBag can accurately identify known structural similarities. We propose that the method which is based on surface rather than fold similarity will be of great advantage for function prediction from distantly related proteins as well as proteins which have evolved by convergent evolution.

METHODS
As a first step, the protein surface is represented by a bag of all its overlapping equal-size (n) surface patches. Surface patches are defined by a central surface residue and its nearest surface residues, identified by the Euclidean distance between their C-alpha atoms. A surface residue is defined if the accessible surface area of its C-alpha atom is higher than 20%, as calculated by DSSP [4]. To cluster the patches, all versus all patches were compared. The distance between any two patches is the minimal root mean square deviation (RMSD), calculated using Kabsch’s algorithm [5], for all possible alignments between the two sets of points. Let n be the number of points in a patch, then the complexity of comparing 2 surface patches is O(n^2), as the time complexity of Kabsch’s algorithm is linear in the number of the points in the set. Then, using the k-means++ algorithm [6], 5,056 (25%) random patches from the training set were clustered to 20, 100, 400 and 800 clusters. The medoid of each cluster was taken to form libraries of varying sizes [7].

Using a similar approach to FragBag [8], each protein was further characterized by a ‘bag-of-surface patches’: a vector representing the number of times each library-patch best approximates a surface patch of a given protein. The similarity between two structures was then measured by the cosine distance of their corresponding vectors (Fig. 1).

RESULTS & CONCLUSIONS
Our training set included 198 representative proteins from [9]. To test the PatchBag approach we used 50 randomly chosen domains from a set of 2,928 sequence-diverse CATH v.2.4 domains [9], with 4,123 patches. Then, the local-fit criterion was used to measure how well each library approximates a protein surface.

The local-fit score of a protein surface was constructed by finding for each surface patch the most similar patch in the library. Then, the average of minimal RMSD values over all patches in all proteins in the test set was calculated to evaluate the local-fit score for each particular library.

Table 1 presents the average PatchBag distances between structure pairs calculated at the different levels of the CATH classification, using a library of size 100. As expected, the average cosine distance grows as the sets become more structurally diverse. Our results clearly demonstrate that the PatchBag approach, which is based solely on the proteins surface, can accurately identify known structural similarities identified based on the protein fold.

REFERENCES

Figure 1 A PatchBag representation of a protein surface. (1) Surface patches are defined by a central surface residue and its n-1 nearest neighbors. Patches were extracted from the training set and clustered to libraries. In this example, the library size =6 and the patch size = 5 (2) For each protein in the test set we extract all (overlapping) surface patches, each patch is matched to the library-patch that best approximates it (in the illustration only 5 patches are shown). (3) The protein surface is represented by the PatchBag vector.

Overall, the local fit, achieved using libraries of different sizes, was 1Å or less. As expected, the larger the library, the better the approximations. The library of size 100 with a local-fit accuracy of 0.88 Å was further chosen for testing.
The central role of post-translational modification in the interactome

Mark N Wass1, Alfonso Valencia2 & Bernard Jacq2

1Structural Biology and Bioinformatics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain and 2Developmental Biology Institute of Marseille-Luminy (IBDML), CNRS and Aix-Marseille University, France. mwass@cnio.es

The functional role of many post-translational modifications (PTMs) is unknown. Performing a structural analysis we demonstrate that PTMs are enriched at interface sites in both human and yeast. There is a strong correlation between presence of PTMs and the number of interaction partners. We additionally observe the potential for extensive PTM crosstalk in these species.

Introduction

Proteomics studies have now identified thousands of post-translational modifications (PTMs) [1] but knowledge of the function of many PTMs is limited. PTMs are known to play a role in the regulation of Protein-Protein interactions (PPIs) through small modules such as SH3 domains. Here we investigate the extent of PTM regulation of PPIs in Homo sapiens, Sacharomyces cerevisiae. Additionally experimental research is beginning to identify examples of crosstalk between PTMs [2-4]. This occurs when two post-translational modifications, often of different types, together regulate a biological process. Such crosstalk generally involves PTMs that share the same site on the protein or that are very close by, with the presence of one of the PTMs being mutually exclusive, although in some examples the sites can be remote [2]. To date PTM crosstalk is probably best characterised for O-linked glycosylation and phosphorylation, although other examples include phosphorylation and methylation [5]. Additionally a recent study has demonstrated extensive crosstalk between phosphorylation and acetylation in Mycoplasma pneumonia [4].

Methods

We used 8 types of PTMs for which there are 150,000 experimentally verified PTMs present in Human and Yeast. Protein-Protein interactions were obtained from Bio-Grid and complexes representing pairs of interacting proteins were identified using BLAST to search the proteins present in Protein Data Bank (PDB) biological units.

Results & Conclusions

We initially mapped the PTMs onto the human and yeast interactomes. This enables an analysis of relationship between PTMs and PPIs. We observe a positive correlation between the presence of PTMs and number of interaction partners for most PTM types, while a negative correlation is found for a few PTM types.

We performed a structural analysis to identify if the PTMs that have a positive correlation with interaction number play a direct role in protein-protein interactions. Using the PDB we identified complexes that represent interactions present in the interactome and mapped the PTMs onto the protein structures. We find that at least three types of PTM are enriched at protein-protein interface sites. We also observe examples where multiple positions within an interface can harbour PTMs. For example Ku a heterodimer involved in the repair of single strand DNA breaks has multiple ubiquitination sites, many of which occur in the interface site of the two proteins (Figure 1). Poly-ubiquitination of Ku is required to remove it from repaired DNA and disrupt the interaction between the two protein subunits [7].

To investigate the extent of PTM crosstalk we first considered the co-occurrence of PTMs on the same protein, which we call internal crosstalk. We observed a high level of co-occurrence for many pairs of PTMs. As PTM crosstalk often occurs when residues are in close proximity we then performed our analysis at the residue level, identifying how frequently PTMs occur at the same position or within 5 residues of each other. We compared the observed frequency with that expected by chance and find that 4 pairs of different PTMs are more frequently in close proximity than expected by chance. Additionally 2 PTM types are more frequently in close proximity with PTMs of the same type than expected by chance. These initial results suggest that PTM crosstalk is widespread and may play important roles in regulation of cellular processes. We observe that proteins where crosstalk occurs have on average more PPIs than those without crosstalk, suggesting that PTM crosstalk plays a role in the regulation of PPIs.

In summary our results demonstrate that post-translational modification plays a central role in the interactome through the regulation of PPIs. Further we observe the potential for widespread PTM crosstalk, which appears to be linked to the regulation of protein-protein interactions.

References

IDENTIFICATION AND VISUALIZATION OF SYSTEMATIC ERRORS IN LOW-RESOLUTION PROTEIN MODELS

Bradley Hintze, Christopher Williams, Daniel Keedy, David Richardson & Jane Richardson
Department of Biochemistry, Duke University. ‘dcrjsr@kinemage.biochem.duke.edu

The field of structural biology provides critical structural details of macromolecules having research utility in a wide variety of disciplines. The utility of a given model is proportional to that model’s accuracy. Solving accurate structures, free of impossible errors, remains an active area of research, especially when working at low resolution. Critical to this goal is identifying errors and creating methods to avoid them in the first place. This research explores exciting new methodologies and new ways to visualize structure validation criteria.

INTRODUCTION
An especially difficult challenge in structural biology is accurate structure determination of macromolecules that yield low-resolution x-ray diffraction data. To eliminate impossible errors one must first identify them and represent them efficiently to people and/or software. Our lab has a long history of dealing with such issues and continues to actively research macromolecular structure validation methods. This research focuses on identifying systematic errors in low-resolution (3-4 Å) models with the ultimate goal of creating tools capable of preventing the introduction of such errors during model building and refinement.

METHODS
A powerful approach in identifying errors in low-resolution PDB models is to compare high/low-resolution pairs. We created a database by parsing the Protein Data Bank (PDB) 90% identity sequence clusters. From each cluster we selected two structures, if present: the structure solved at the lowest resolution, ≤2.8 Å, and the other structure solved at highest resolution, ≤2.0 Å, at a later date. In this way the high-resolution model acts as a reference standard to the low-resolution model, which typically has many more errors.

MolProbity Compare is a program we created that makes a pairwise comparison of validation criteria for structure pairs on a per-residue basis (see figure). The output is a scrollable table with each column representing equivalent residues, one from each model, and each row reporting different validation criteria. Since the columns are in sequential order this visualization allows the user to quickly identify regions that are better in one model compared to the other. This tool has broad utility in the field and will be available on our MolProbity site.

CABLAM is an exciting new program currently being developed in the lab that uses virtual dihedral angles to characterize local conformations of protein backbone. Two purely Cα-based dihedrals can identify secondary structure reliably even at low resolution, while a third dimension that relates adjacent C-O directions can diagnose local errors.

RESULTS & CONCLUSIONS
We used the MolProbity Compare analysis to create a distribution of the number of outliers per residue. This analysis allows us define thresholds and distinctive patterns of errors to identify an out-of-register region of sequence in a low-resolution model. We believe that by adding per-residue real-space residuals we will further increase the sensitivity of this analysis.

CABLAM analysis can identify backbone regions that differ conformationally between two otherwise similar models. We can readily identify incorrect carbonyl flips between two models using CABLAM and Ramachandran analysis; flips occur where CABLAM shows similarity in a region but two adjacent residues differ significantly in Ramachandran values. This analysis will be expanded to use C-O dihedrals to explore possible automated ways of identifying incorrectly fit peptides in a single structure.

The methods presented here can identify errors within low-resolution atomic models. This is a first step in the larger goal of developing tools for the crystallography community that can correct or avoid such errors, leading to more accurate structures for large, biologically significant macromolecules and complexes.

REFERENCES
RNA 3D HUB, WEBFR3D, AND R3DALIGN - DATABASES AND WEB APPLICATIONS FOR RNA STRUCTURAL BIOINFORMATICS

Anton I. Petrov1, Ryan Rahr2, Blake Sweeney2, Craig L. Zirbel1 & Neocles B. Leonitis4
Depts. of Biological Sciences1, Mathematics and Statistics1, and Chemistry2, Bowling Green State University.
Dept. of Mathematics and Statistics, Ohio Northern University. leontis@bgsu.edu

We present a new online resource called RNA 3D Hub, which houses a representative collection of RNA 3D motifs called RNA 3D Motif Atlas and non-redundant sets of RNA-containing 3D structures. In addition, we present WebFR3D and R3DAlign, two web applications for searching and superimposing RNA 3D structures. RNA 3D Hub, WebFR3D, and R3DAlign provide an online suite of tools for studying and exploring RNA 3D structure.

INTRODUCTION

Many internal and hairpin loops, which are usually drawn as unstructured in RNA secondary structure diagrams, in fact, form intricate structural motifs stabilized by non-Watson-Crick interactions. These RNA 3D motifs are often recurrent and are essential for many biological functions and RNA folding.

Although more and more new databases and web applications related to RNA structural bioinformatics are released each year, there is still no comprehensive resource for RNA structural motifs and no common approach has been developed to address the redundancy inherent in the RNA structural data.

Thus, it is desirable to create a comprehensive collection of RNA 3D motifs in order to advance the RNA 2D and 3D structure prediction and ncRNA discovery methods, to interpret mutations that affect ncRNAs, and to guide experimental functional studies.

Here we present a new online resource called RNA 3D Hub (http://rna.bgsu.edu/rna3dhub), which houses RNA 3D motifs extracted from all atomic-resolution RNA 3D structures deposited in PDB/NDB, and the RNA 3D Motif Atlas, a representative collection of RNA 3D motifs. RNA 3D Hub also hosts non-redundant sets of RNA-containing 3D structures and structural annotations of various pairwise interactions found in 3D structures, such as basepairing and stacking.

METHODS

We use an automated procedure to produce non-redundant lists of RNA 3D structures based on sequence similarity, 3D superpositions, and structure quality considerations. Next, we extract RNA 3D motif instances from the current non-redundant list using FR3D [1], a program for symbolic and geometric searching of RNA 3D structures. Finally, we use a clustering approach to obtain a representative collection of RNA 3D motifs. All these data are stored in RNA 3D Hub and can be accessed online. Unique and stable ids are assigned to all motifs, to all motif instances, and to all non-redundant equivalence classes of structure files. RNA 3D Hub is updated automatically on a regular schedule, and a versioning system is implemented to provide independent access to data snapshots. RNA 3D Hub also provides a rich user interface to allow for efficient data visualization (Figure 1).

RESULTS & CONCLUSIONS

RNA 3D Hub complements a suite of web applications for searching and superposing RNA 3D structures that is also presented. WebFR3D [2] (http://rna.bgsu.edu/webfr3d) is the online version of FR3D [1], a program for symbolic and geometric searching of RNA 3D structures. Using WebFR3D users can search for RNA 3D motifs in individual PDB files or in non-redundant sets of RNA structures. R3DAlign (http://rna.bgsu.edu/r3dalign) is a program for global pairwise structural alignment of RNA 3D structures using local superpositions. It is available both as a standalone program and a web application [3]. While WebFR3D is designed to search for relatively small motifs, R3DAlign can be used to align complete RNA 3D structures while allowing for differences in the global structure of the molecules.

REFERENCES


![Figure 1. RNA 3D Motif Atlas entry for the sarcin/ricin internal loop motif. The annotations of base pairing, base stacking, and base-backbone interactions that characterize the motif are listed in the table (top). Users are able to interactively view and superimpose motif instances and to explore their structural context (bottom left). Mutual similarity between all motif instances is visualized as a heatmap (bottom right) with similar structures colored in red and yellow.](http://rna.bgsu.edu/RNA3DHub/05.png)
PROTEIN STRUCTURE ALIGNMENT BEYOND SPATIAL PROXIMITY
Sheng Wang¹, Jianzhu Ma¹, Jian Peng¹ & Jinbo Xu¹²
Toyota Technological Institute at Chicago¹,²; jinboxa@gmail.com

The problem of automatically constructing an accurate protein structure alignment still remains challenging especially when proteins to be aligned are distantly-related. We present a novel method, DeepAlign, which aligns two protein structures using not only spatial proximity of equivalent residues (after rigid-body superposition), but also evolutionary distance and hydrogen bonding similarity. Experiments show that DeepAlign can generate structure alignments highly consistent with human-curated alignments, outperforming others greatly.

INTRODUCTION
The alignment accuracy of existing protein structure alignment algorithms is still low when judged by manually-curated structure alignments, especially when distantly but functionally related proteins are aligned. The performance of an alignment algorithm mainly depends on two components: a scoring function measuring protein structure similarity and a search algorithm optimizing the scoring function. It is very challenging to design a scoring function to capture all the (implicit and explicit) criteria used by human experts, who align protein structures using not only geometric information, but also evolutionary and functional information. It is also not trivial to optimize a scoring function. Most search algorithms can only solve scoring functions to suboptimal and thus, generate a suboptimal alignment, which may impact alignment accuracy. Our method DeepAlign makes use of a novel scoring function and also an efficient search algorithm. Experiments show that DeepAlign can generate significantly better alignments than others especially in aligning distantly-related proteins.

METHODS
Scoring function. DeepAlign uses the following scoring function to measure the similarity of two aligned residues i and j.

\[
\text{Score}(i, j) = \max(0, \text{BLOSUM}(i, j) + \text{CLESUM}(i, j))
\]

where CLESUM is the local structure substitution matrix and BLOSUM is the amino acid substitution matrix; \(v(i, j)\) measures the hydrogen-bonding similarity using three vectors; and \(d(i, j)\) measures the spatial proximity of two aligned residues (after rigid-body superposition) using TMscore. This scoring function is derived from the observation that, despite two homologous proteins share a similar overall conformation, their structures exhibit very high local flexibility due to evolutionary events happening at the sequence level. This kind of local flexibility cannot be accurately described by spatial proximity of aligned residues. Instead, the evolutionary distance, measured by amino acid or local substructure substitution matrix, as well as the hydrogen bonding similarity, may be more appropriate.

Search algorithms. DeepAlign starts from similar substructures (determined by amino acid and local substructure mutation matrices) to identify a small subset of equivalent residues. Then DeepAlign iteratively optimizes the scoring function and generates the best alignment using the procedure similar to TMalign. Since the similar substructures are determined by evolutionary distance instead of spatial proximity, DeepAlign aligns evolutionary-related residues and avoids many false positives (which are purely determined by spatial proximity). As such, DeepAlign can quickly generate alignments that make biological sense.

RESULTS & CONCLUSIONS
DeepAlign can generate structure alignments highly consistent with human-curated alignments on three benchmarks CCD¹, MALIDUP² and MALISAM³. DeepAlign greatly outperforms other tools especially when proteins under consideration (i.e., those in MALISAM) are distantly-related (see Table 1).

Table 1. Alignment accuracy of DeepAlign, DALI, MATT and TMalign on 3 human-curated benchmarks CCD, MALIDUP and MALISAM. The human-curated alignments are used as the ground truth.

<table>
<thead>
<tr>
<th>Method</th>
<th>CDD</th>
<th>MALIDUP</th>
<th>MALISAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepAlign</td>
<td>93.8</td>
<td>92.0</td>
<td>77.5</td>
</tr>
<tr>
<td>DALI</td>
<td>92.8</td>
<td>86.4</td>
<td>67.7</td>
</tr>
<tr>
<td>MATT</td>
<td>91.4</td>
<td>79.8</td>
<td>51.7</td>
</tr>
<tr>
<td>TMalign</td>
<td>85.6</td>
<td>81.0</td>
<td>53.7</td>
</tr>
</tbody>
</table>

DeepAlign can also classify protein domains much more consistently with SCOP, a human-curated protein classification database (see Figure 1).

![DeepAlign ROC curve](image)

DeepAlign can generate M (M=10) alternative alignments for a protein pair. As shown in Table 2, even if only two alternative alignments are used, DeepAlign can greatly improve alignment accuracy.

Table 2. The alignment accuracy of the best among the top M alternative alignments generated by DeepAlign on three manually-curated benchmarks CCD, MALIDUP and MALISAM. The human-curated alignments are used as the ground truth.

<table>
<thead>
<tr>
<th>DeepAlign Top M</th>
<th>CDD</th>
<th>MALIDUP</th>
<th>MALISAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top1</td>
<td>95.8</td>
<td>92.0</td>
<td>77.5</td>
</tr>
<tr>
<td>Top2</td>
<td>95.1</td>
<td>92.7</td>
<td>87.6</td>
</tr>
<tr>
<td>Top3</td>
<td>95.2</td>
<td>93.1</td>
<td>88.0</td>
</tr>
<tr>
<td>Top4</td>
<td>95.3</td>
<td>93.1</td>
<td>88.9</td>
</tr>
<tr>
<td>Top5</td>
<td>95.3</td>
<td>93.1</td>
<td>89.7</td>
</tr>
<tr>
<td>Top10</td>
<td>95.3</td>
<td>93.1</td>
<td>90.0</td>
</tr>
</tbody>
</table>

These results imply that to generate biologically meaningful protein structure alignments, we shall also consider evolutionary relationship in addition to geometric similarity.

REFERENCES
AN UPDATED COMPARISON OF METHODS FOR REMOTE HOMOLOGY DETECTION

Hai Fang1, David de Lima Moraís2 & Julian Gough2*
Dept. of Computer Science1, University of Bristol. *goughh@cs.bris.ac.uk

In the last ten years, coming to fruition recently, there have been some major advances in sequence comparison methods, the two most important being: profile-profile methods and heuristic acceleration of hidden Markov model scoring. Ten years ago we conducted an independent benchmark which has provided users with valuable guidance and indicated to developers where there is potential for further gain. Here we present a timely update.

INTRODUCTION

Sequence homology detection is one of the central problems in bioinformatics. It has been used in one form or another by most scientists in the field of molecular biology, and by almost all of those in the sub-field of protein structure. Few advances (but highly significant ones) have been made in the latter half of the last decade, despite the well-advanced state of the art ten years ago. An independent comparison of available software is now needed to inform scientists which methods they should be using and to apply pressure to the community to move forward from using older methods which have become habitual. Good independent benchmarks are crucial for driving the technology developers to improve and suggest where the gains can be made.

We present here a timely update, a decade on from our 2002 paper1 “A comparison of hidden Markov model procedures for remote homology detection”. This 2002 paper continues to be influential, despite the well documented development of methods in the last decade, so it is important for us to provide the replacement, and to make efforts to make the community aware of the new results.

METHODS

The benchmark in this work is based on the SCOP2 database of known structural domains classified into superfamilies for which there is evidence for a common evolutionary ancestor. The data set and rules are in essence the same as in the 2002 paper, but have been updated to the current 1.75 version of SCOP. In brief, sequences of SCOP domains filtered to 95% sequence identity are scored with each method in an all-against-all test ignoring self-hits; results are marked as true/false based on the SCOP classification. From best to worst score, a hit to the same superfamly is counted true, and a hit to a different fold is counted false. There is a list of exceptions to the SCOP classification such as TIM barrels and Rossmann domains (these mostly appear as comments on the SCOP website).

The methods we tested include profile-profile comparison methods—one of the major advances of the last decade— as well as profile-sequence and sequence-sequence methods. Described above is the benchmark for sequence-sequence comparisons, but for profile-based methods we ran benchmarks both with automatically produced models seeded with each sequence, and with an example of state-of-the-art curated models (from the SUPERFAMILY database).

We conducted the benchmarks both on sequences of domains directly, and also with the domain sequences embedded in longer sequences emulating a multi-domain context for local scoring. Some methods, or stages of methods, require a large background sequence database: for this we used the same version of UniProt90 in all cases.

The rough computational cost of all methods was also tested, adding an important context to the comparison.

RESULTS & CONCLUSIONS

The main sources of software compared come from BLAST, HMMER, SAM and HHSuite but some others are included. These all have many components which were assessed independently and in concert.

Pair-wise sequence comparison methods perform poorly compared to profile methods, however they are still important because they are needed as the first step of any iterative procedure for creating multiple sequence alignments for building profiles. We compared: BLAST, PHMMER, CS-BLAST and HHblits (sequence mode). In terms of pure profile-profile methods we compared HHsearch, PRC and HHblits (profile mode).

Profile-sequence methods include model building and model scoring; we compared most possible combinations of HMMER2, HMMER3 (with/without heuristic speed-ups) and SAM components and some combinations from HHblits and PSI-BLAST.

The main packages have an iterative procedure that makes use of sequence-sequence search, model building and model scoring on a background sequence database to build a profile from an initial seed sequence: PSI-BLAST, JackHMMER, SAM-T99, SAM-T2K, SAM-T6K and HHblits (iterative mode). We compare these, including the effect of different numbers of iterations.

REFERENCES


FIGURE 1. A few selected results showing the difference between sequence-sequence, sequence-profile and profile-profile methods.

There are many results, some of which require detailed interpretation, but a few of the most basic are shown in Figure 1. Of the many lessons learned, there is space here for three: HMMER3 heuristic speed increases, iterative JackHMMER and ease of use mean there is no longer any justification for using PSI-BLAST; HHblits makes profile-profile possible on genomes or large sequence databases; components of HMMER3 and HHblits can still be much improved.
We design a novel statistical potential EPAD by making use of protein evolutionary information, which has not been used by currently popular statistical potentials (e.g., DOPE and DFIRE). EPAD is unique in that it has different energy profiles for the same type of atom pairs, depending on their sequence positions. EPAD outperforms several popular statistical potentials in both decoy discrimination and ab initio folding.

**RESULTS & CONCLUSIONS**

Fig. 1 shows the EPAD and DOPE potentials as a function of inter-atom distance for the atom pair ALA Cα and LEU Cα in 3 different positions of protein 1grp. DOPE has the same energy profile regardless of the sequence positions. The DOPE potential is favourable only when the atomic distance is 5-Å. In contrast, EPAD has a different energy profile at each position. The EPAD potential is favorable when the distance of this atom pair is close to the native.

**FIG 1.** Distance dependence of the DOPE and EPAD potentials for the same atom pair in 3 different positions of protein 1grp. The figure legend shows the native atomic distances at the 3 positions.

Table 1 shows that EPAD greatly outperforms several popular potentials including DOPE, DFIRE, OPUS and RW in discriminating decoys in the Rosetta set. The same trend is also observed when several other decoy sets are tested including the CASP9 models, the I-TASSER dataset and the CASP5-8 dataset.

**TABLE 1.** EPAD greatly outperforms several popular potentials on the Rosetta decoy sets. Metrics used: the number of correctly identified natives, Z-score of the native energy, the average model quality, measured by GDT, of the first-ranked decoys and the Pearson correlation coefficient between the energy and GDT.

<table>
<thead>
<tr>
<th>Method</th>
<th>#natives identified</th>
<th>11</th>
<th>12</th>
<th>6</th>
<th>first-ranked GDT</th>
<th>Pearson CC</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPAD</td>
<td>33</td>
<td>4.59</td>
<td>47.0</td>
<td>47.8</td>
<td>45.9</td>
<td>-0.64</td>
<td>-0.24</td>
</tr>
<tr>
<td>DOPE</td>
<td>12</td>
<td>4.49</td>
<td>47.0</td>
<td>47.8</td>
<td>45.9</td>
<td>-0.64</td>
<td>-0.24</td>
</tr>
<tr>
<td>DFIRE</td>
<td>12</td>
<td>4.59</td>
<td>47.0</td>
<td>47.8</td>
<td>45.9</td>
<td>-0.64</td>
<td>-0.24</td>
</tr>
<tr>
<td>OPUS</td>
<td>6</td>
<td>4.59</td>
<td>47.0</td>
<td>47.8</td>
<td>45.9</td>
<td>-0.64</td>
<td>-0.24</td>
</tr>
<tr>
<td>RW</td>
<td>6</td>
<td>4.59</td>
<td>47.0</td>
<td>47.8</td>
<td>45.9</td>
<td>-0.64</td>
<td>-0.24</td>
</tr>
</tbody>
</table>

**REFERENCES**

Three dimensional structure elucidation of acyl-CoA-binding proteins from Tung Oil and Arabidopsis thaliana using Robetta full-chain structure prediction server

Abul H. J. Ullah1,2 and Jay Shockey1

1Southern Regional Research Center, ARS, USDA, New Orleans, LA 70124
2abul.ullah@ars.usda.gov

INTRODUCTION
Acyl-CoA-binding proteins (ACBPs) are a family of proteins exhibiting conservation in an acyl-CoA-binding domain which binds Acyl-CoA esters with high specificity and affinity. ACBPs are ubiquitous in animals, plants, fungi and protists but not in prokaryotes and archaea. Plant 10-kDa ACBPs were first identified from Brassica napus and Arabidopsis thaliana.

METHODS
The sequences of three ACBPs of varying lengths from Vernicia fordii (Tung oil plant) and one short ACBP from Arabidopsis thaliana were submitted to Robetta full chain protein structure prediction server (http://robetta.bakerlab.org) for elucidation of three-dimensional structure and analysis. For structure prediction, sequence submitted to the server was parsed into putative domains and structural models were generated using either ROSETTA fragment insertion method. Domains without a detectable PDB homolog were modelled with the Rosetta de novo protocol. Comparative models were built from Parent PDBs detected by UW-PDB-BLAST or HHSEARCH and aligned by various methods which include HHSEARCH, Compass, and Promals. Loop regions were assembled from fragments and optimized to fit the aligned template structure. The procedure is fully automated.

RESULTS & CONCLUSIONS
While V. fordii ACBP3A first domain (1-153) is composed of 5 α helices (Figure 1), V. fordii ACBP3B first domain (1-203) showed 7 helices (Figure 2). V. fordii ACBP6 is a much shorter protein (1-91), which when modelled showed 4 helices (Figure 2). We also modelled the Tung Oil Plant ACBP6, which is a much shorter protein (Figure 3). This also depicted a 4 helices bundle pattern. For comparison, we modelled a shorter version of Arabidopsis thaliana, a cruciferous plant ACBP3, which was made short (93 aa) from the full-length (362 aa) by deleting the last 269 residues from the c-terminus; this shortened protein also exhibited a 4 helices bundle profile – a characteristic of ACBP. When we compared the plant ACBPs with the human ACBP, which was modelled from X-ray diffraction data, we visualize again a 4 helices bundle profile. Therefore, it is fair to assume that a minimum of 4 helices bundle is required for creating the binding site for Acyl-CoA ester in ACBPs. The Robetta full chain structure prediction server is thus providing the scientific community an alternative way to deduce three-dimensional structure of naturally occurring proteins. Our laboratory is taking the full advantage of this service to refine the structure of V. fordii ACBPs to optimize TAG synthesis in Tung Oil tree. We are presently measuring the binding constant (Kd) of these ACBPs with [14C]oleoyl-CoA ester. The site-directed mutagenesis approach could be used to disrupt any of these helices to map the binding site.

REFERENCES
FROM CHEMICAL TO SYSTEMS BIOLOGY: HOW NETWORK PHARMACOLOGY CAN CONTRIBUTE TO IT?
Karine Audouze, Tudor Oprea, Soren Brunak, Olivier Taboureau

1 Center for biological sequences analysis, Dept of Systems Biology, Technical University of Denmark
2 Division of Biocomputing, Dept of Biochemistry and Molecular Biology, University of New Mexico School of Medicine, Albuquerque, NM, USA otab@cbs.dtu.dk

With the necessity to investigate alternative approaches and emerging technologies in order to increase drug efficacy and to reduce their adverse effects, systems pharmacology offers a novel way of approaching drug discovery by considering the effect of a molecule and protein’s function in a global physiological environment. By studying drug action across multiple scales of complexity, from molecular to cellular and tissue levels, network-based computational methods have the potential to improve our understanding on the impact of chemicals in human health.

INTRODUCTION
The old drug discovery paradigm, i.e. one target one drug, has been traditionally the driving force for developments in biomedicine. Recent advances in chemical biology and systems biology have shown that most drugs interact with multiple targets and the pharmacological profile of a drug is not as reductionist as we believed. Based on the available information related to drugs, it is estimated that a drug interacts to 30 000 protein that we translate the integration here is the protein. This is through the collected, integration of all the be mapped to protein (PharmGKB) and KEGG and can be collected from large interactions from both humans and model organisms relation to the chemical entities. Protein side, a plethora of data can data from literature i.e. ChEMBL. From the biological side, a plethora of data can be of potential interest in relation to the chemical entities. Protein-protein interactions from both humans and model organisms can be collected from large compendiums (BIND, MINT, HPRD). Disease-gene associations (OMIM, and GeneCards), pathways information (Reactome and KEGG), gene expressions in tissues (Human Protein Atlas and GNF), genetic variations and clinical effect (PharmGKB), are also biological information that can be mapped to proteins and genes. Once the data collected, integration of all these information was performed. The central feature considered in such integration here is the protein. This is through the protein that we translate the information from chemical effects to biological phenotypes.

METHODS
Data gathering and integration from the chemistry side and biology side were considered in this study. Since few years, a variety of public domain chemical bioactivity resources have been developed and are accessible for the scientific community. They can be categorized as databases providing access to high-throughput screening (HTS) experiments i.e. PubChem or databases gathering chemical bioactivity data from literature i.e. ChEMBL. From the biological side, a plethora of data can be of potential interest in relation to the chemical entities. Protein-protein interactions between both humans and model organisms can be collected from large compendiums (BIND, MINT, HPRD). Disease-gene associations (OMIM, and GeneCards), pathways information (Reactome and KEGG), gene expressions in tissues (Human Protein Atlas and GNF), genetic variations and clinical effect (PharmGKB), are also biological information that can be mapped to proteins and genes. Once the data collected, integration of all this information was performed. The central feature considered in such integration here is the protein. This is through the protein that we translate the information from chemical effects to biological phenotypes.

RESULTS & CONCLUSIONS
We have developed a chemogenomics platform of more than one million of unique compounds with bioactivities on more than 30 000 human proteins. Based on the chemical structures, all compounds were encoded to diverse structural fingerprints (2D and 3D) providing the possibility to suggest new bioactivities for a compound of interest. From the protein, biological information can be directly mapped to it. In addition, we studied phenotypes through protein complexes, gathered from the protein-protein interactions networks. In this case we considered that proteins that interact each other’s have a tendency to be involved in the same phenotypes and processes. Such protein complexes can enrich the information regarding phenotypes, diseases and suggest new potential targets and off-targets perturbed by a compound. The concept is summarizing in Figure 1.

FIGURE 1. Based on a chemogenomic platform, the figure represents the different aspects that can be considered to evaluate the chemical effect on systems biology.

In conclusion, it is now possible to integrate a large amount of biological data, from molecular level to systems and clinical levels, to a chemical (and its structure), providing a deeper understanding of the drug effect in drug discovery and also the potential impact in human health.

REFERENCES

3DSIG: STRUCTURAL BIOINFORMATICS AND COMPUTATIONAL BIOPHYSICS
AUDOUZE ET AL. ID 42 LONG BEACH, JULY 13-14 2012
PROTEIN THREADING WITH SPARSE SEQUENCE PROFILES

Jianzhu Ma1, Jian Peng1, Sheng Wang2 & Jinbo Xu1,

Toyota Technological Institute at Chicago*, jinboxu@gmail.com

Profile-based methods for template-based protein modeling usually fail when proteins under consideration have sparse sequence profiles. To deal with this, we present a novel protein threading method, CNFpred, which employs a probabilistic graphical model Conditional Neural Field (CNF) to align two proteins by using a nonlinear scoring function to combine a variety of sequence and structure information. Experiments show that CNFpred generates significantly better alignments than other profile and threading methods especially for distantly-related proteins with sparse sequence profiles.

INTRODUCTION

Template-based protein modeling (TM) methods including homology modeling and protein threading still suffer from two limitations. The first is that these methods heavily depend on sequence profiles. Although sequence profiles are very powerful, they fail when a protein has a very sparse sequence profile. A protein with a sparse sequence profile has very few non-redundant sequence homologs in the NR database and thus it is more challenging to predict the structure of such a protein unless it has a close template in PDB. The second is that these methods use a linear alignment scoring function to align two proteins. A linear function cannot deal well with correlation among protein features, although many protein features are indeed correlated. To go beyond these limitations, our protein threading method CNFpred introduces several unique features and departs significantly from existing methods.

METHODS

CNFpred is built upon a recently-developed probabilistic graphical model Conditional Neural Field, which integrates the strength of both Conditional Random Fields (CRFs) and neural networks. CNFpred has the following special features.

(1) CNFpred explicitly accounts for correlations among protein features by using a nonlinear scoring function to combine a variety of sequence and structure information.

(2) When a protein has a sparse sequence profile, CNFpred relies more on structural information since the sequence profile does not contain sufficient information; otherwise it relies more on information in a sequence profile. The structure information includes the 3-class and 8-class secondary structure, the 3-class solvent accessibility and also the structure environment.

(3) CNFpred uses neighborhood information to estimate how likely two residues shall be aligned. The neighborhood information includes sequence profile, secondary structure and solvent accessibility in a small window (size 11) centered at the residues to be aligned. Neighborhood information is especially useful to the weakly similar regions and gap opening positions.

(4) For the disordered regions, CNFpred uses only sequence information since structure information is unreliable. For non-disordered regions, CNFpred uses both sequence and structure information.

(5) Unlike many other methods that use an affine gap penalty, CNFpred uses both position-specific and context-specific gap penalty. The position-specific gap penalty is derived from the alignment of the sequence homologs of a given protein while the context-specific penalty is based upon amino acid identity, hydrophathy index, secondary structure and solvent accessibility. When a protein has a sparse sequence profile, CNFpred relies more on context-specific gap penalty; otherwise on the position-specific penalty.

RESULTS & CONCLUSIONS

We evaluate CNFpred using 4 benchmarks (see Table 1) in terms of the reference-dependent alignment accuracy and the quality of the resulting 3D models. The quality of a 3D model is evaluated by TMscore, ranging from 0 (the worst) to 1 (the perfect).

(1) In terms of ref-dependent alignment accuracy CNFpred is >10% better than the best profile method HHpred regardless of the benchmarks and the structure alignment tools used to generate reference alignments.

(2) To evaluate the quality of the resulting 3D models, given a protein pair we build a 3D model using MODELLER for the target protein based upon its alignment to the template. As shown in Table 1, CNFpred obtains much better 3D models than HHpred, MUSTER and BThreader regardless of the benchmarks, outperforming the best profile method HHpred by 7-20%.

Table 1. The accumulative model quality, measured by TMscore, on the four benchmarks: In-House, MUSTER, SALIGN and ProSup. Bold indicates the best performance.

<table>
<thead>
<tr>
<th>Methods</th>
<th>In-House</th>
<th>MUSTER</th>
<th>SALIGN</th>
<th>ProSup</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHpred</td>
<td>1522.77</td>
<td>142.00</td>
<td>121.83</td>
<td>56.44</td>
</tr>
<tr>
<td>MUSTER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BThreader</td>
<td>1537.89</td>
<td>143.95</td>
<td>132.85</td>
<td>66.77</td>
</tr>
<tr>
<td>CNFpred</td>
<td>1692.17</td>
<td>152.14</td>
<td>134.50</td>
<td>67.34</td>
</tr>
</tbody>
</table>

(3) To evaluate the modeling performance, we use CNFpred and HHpred to predict the 3D structure for a set of 1000 target proteins randomly chosen from PDB25. All the ~6000 proteins in PDB25 are used as the templates. As shown in the below Figure, CNFpred greatly outperforms HHpred especially when the target protein does not have a close template. Each point in the figure represents two models of a single target. One is built by HHpred and the other by CNFpred. A point above the diagonal line indicates that CNFpred generates a better 3D model for the target.

As shown in the figure, the targets with HHpred TMscore<0.4 usually have sparse sequence profiles and thus HHpred does not work well for them. By contrast, CNFpred can generate much better 3D models for many of them.
**SYSTEMATIC SURVEY OF INTERTWINED HOMOMERS**

Stephen MacKinnon, Michael Garton, Anatoly Malevanets, & Shoshana J Wodak

2. Dept. of Biochemistry and Molecular Genetics, University of Toronto, Toronto, Canada.
3. The Centre for Biomolecular Sciences, University of Nottingham, UK.

We systematically survey the Protein DataBank to catalog all known instances of intertwined homomeric proteins. The properties of intertwined interfaces are analyzed and their role in stabilizing quaternary homomer assemblies is documented. We find that intertwining geometries are well conserved, even at low sequence identity (<40%), and hence likely dictated by the protein topology. Our study furthermore offers new insights into how intertwining may modulate protein dynamics and function.

**INTRODUCTION**

A large fraction of all proteins assemble with identical counterparts to form homo-oligomeric structures, also known as homomers. The vast majority homomers adopt cyclic, tetrahedral, or dihedral symmetry. Homomers with extensive inter-subunit interactions sometimes described as *intertwined* or *swapped*. In these cases, two polypeptide chains appear entangled and form a single structural entity. An example intertwined cyclic trimer is illustrated in Figure 1. Some intertwined proteins are known to assemble via 3D Domain Swapping. Domain Swapping is the interconversion between a monomeric protein and a homomeric state, formed by exchanging intramolecular contacts for their intermolecular equivalents. 3D domain swapping has been observed in several protein systems known to misfold and lead to disease in humans. Previously, intertwined proteins and 3D domain swapping were only described qualitatively and studies seeking to understand their general behavior were limited to manually-collected, small subsets of protein structures. We present an automated analysis of intertwined associations in homomeric proteins of known structure. Our aim is to gain mechanistic insights into the formation and role (physical and functional) of these complex association modes, where the high degree of inter-subunit contacts likely reflects that fact that folding and conformational changes are coupled to subunit assembly. To this end, we investigate the properties of intertwined interfaces as well as their role in stabilizing homomer quaternary structures. We also evaluate the extent to which intertwined geometries are conserved across families of related proteins, which provides information on the respective roles of the protein topology and sequence in dictating these geometries. Finally preliminary results are presented on how the formation of intertwined associations can modify the dynamic properties of the protein, and how this modification may modulate protein function.

**METHODS**

We formally define ‘intertwined proteins’ as any symmetric homomer that forms intermolecular, non-contiguous structural domains (NCSDs). We then use domain-partitioning algorithms to decompose entire homomorphic assemblies into NCSDs, simultaneously identifying intertwined proteins and their ‘swapped’ residues. Each NCSD is composed of the non-swapped residues from one subunit and the swapped residues of neighboring polypeptide chain. Individual intertwined associations, or ‘swaps’, are classified according to their propagation: dimeric, a reciprocal swap between two subunits; oligomeric, a circular arrangement of swaps between three or more subunits; or polymeric, a swap propagating indefinitely across a crystal lattice. Associations without apparent symmetry and those with interfaces smaller than 370Å² are excluded from this investigation. Applying this procedure, we survey the Protein Databank (PDB) for intertwined associations. We consider all structures solved by NMR spectroscopy and those solved by X-Ray crystallography with 2.5Å or better resolution.

**RESULTS & CONCLUSIONS**

Upon our initial survey of the Protein Databank, we find that intertwined proteins are remarkably common. Intertwined associations are observed in approximately 25% of all homomers encountered. While common, the degree of intertwining is quite variable, ranging from subtle to extensively intertwined, with some involving the exchange of entire structure domains. We find that the subunit interfaces of intertwined proteins are more tightly packed and have a higher fraction of non-polar atoms than their non-swapped counterparts. These properties are often associated with stable and biologically relevant inter-subunit associations. We report the biases in amino acid and secondary structure compositions for these systems. We discuss the physical and biological roles of intertwined associations. We also present the evolutionary behavior of intertwined proteins, the relationship between intertwined associations and quaternary structure prediction, the role of swapping in dihedral homomers, and a new web tool designed to explore and visualize intertwined proteins. Finally, we present the preliminary results of a new study, investigating the role of conformational disorder in the homomer assembly and function.

**REFERENCES**


![Figure 1](image.png)
**INTRODUCTION**
Electron-transfer mechanisms catalyzed by metal-binding oxidoreductases (Enzyme Commission class 1; EC 1) enable the biogeochemical cycling of H, C, N, O, S, and P\(^2\). The evolutionary distance between these enzymes requires developing specialized search methods to detect relationships. The levels of sequence similarity are often too low to perform this task. One approach to otherwise relate distant proteins is structural comparison. We use the oxidoreductase structures found in the Protein Data Bank (PDB)\(^3\) to propose methods for detecting distant relationships between proteins of this class.

**METHODS**
The TrAnsFuSE project\(^4\) produced a gold standard set of 158 Swiss-Prot\(^5\) sequences containing 35 InterPro\(^6\) domains and matching 976 PDB structures. From this dataset, 138 unique transition-metal containing structural domains were selected as representatives of the structural variety. Each group of protein structures was examined and the best resolution structure, containing all the necessary metal ligands, was selected (Figure 1). All-against-all structure comparisons were performed using the TopMatch algorithm\(^7\) with standard parameters. For each pair the highest ranked alignment was used to calculate the pairwise structural distance\(^8\) between the proteins. The result is a complete network of the relationships between structural protein domains. We also built a phylogenetic tree to trace the patterns in the evolution of structures. Note that we picked the ferredoxin domain to root the tree since it is considered to be one of the oldest electron transfer catalyzing folds.

**RESULTS & CONCLUSIONS**
The network topology is based solely on structural similarities, but highlights clusters of proteins containing similar ligands and equivalent InterPro annotations. Removing from the network all edges where the asymmetric similarity (coverage of query and target structures) is lower than 43\% (highest value to yield no isolated nodes) places the small ferredoxin domain from 1gte\(^9\) chain A (Figure 1) as the most central network node. This result is expected since the ferredoxin fold is small accounting for lower structural distances from other folds. However, we also suspect that its centrality is related to its evolutionary age. The phylogenetic tree, built using structural distances, exhibits interesting characteristics. The tree has two major clades containing mostly alpha-helical OR mostly beta-sheet structures. The tree also confirms that structure comparison is able to relate proteins of similar shapes and functions (e.g. different peroxidases) even when the sequence similarity is very low.
**INTRODUCTION**

Protein modeling is widely used in life science research to build models for proteins, where no experimental structures are available. However, depending on the specific target protein and the applied modeling approach, the accuracy of computational models may vary significantly. CAMEO assesses prediction accuracy on a weekly basis and thus helps methods developers by continuously benchmarking new developments, and users of models to select the most suitable tool for a given modeling problem.

**METHODS**

Sequences of soon to be released entries are retrieved from the PDB and submitted to all registered servers. After the corresponding coordinates have been released by the PDB, CAMEO analyzes the submissions and evaluates their accuracy in comparison to the reference structure (target). A variety of different scores are used to measure the deviations between the models and the targets, including GDT-HA, RMSD, structural coverage of target sequence, local Distance Difference Test (DDT). Additionally, CAMEO evaluates the reliability of the model confidence values (“Model B-Factors”).

**RESULTS & CONCLUSIONS**

Since the accuracy requirements for different scientific applications vary, CAMEO offers a variety of scores assessing different aspects of a prediction (coverage, global accuracy, local accuracy, completeness, Cc and all atom based measures etc.) to reflect these requirements (Figure 1). Notably, the difference in accuracy between different target proteins, or even segments within a single protein, are much larger than the differences observed between different methods for same targets proteins. Therefore, in order for a model to be useful in practical applications, reliable model confidence scores are crucial for users of models to judge if a specific model is sufficient for the planned purpose. CAMEO therefore not only assesses the accuracy of the prediction coordinates, but also if the confidence values assigned to these predictions are reliable.

CAMEO evaluates pre-diction accuracy (not post-dictions), and hence provides an independent blind benchmark of the performance of new algorithms, and the assessment data on a given method can thus be used for publications. We are currently working on the next version of CAMEO, which will include the assessment of the predicted quaternary structure (i.e. oligomeric state) of the predictions, as well as binding sites for ligands and cofactors.

CAMEO forms part of the Protein Model Portal (http://www.proteinfamily.org/) of the Protein Structure Initiative Structural Biology Knowledgebase (http://sbkb.org/).

CAMEO is partially supported by the SIB Swiss Institute of Bioinformatics and by grant 5U01 GM093324-02 from NIGMS for the Protein Structure Initiative Structural Biology Knowledgebase.

**REFERENCES**

INTRODUCTION

The study of evolutionary rates is a central issue to understand the mechanisms underlying protein molecular evolution. Several factors have been associated to the modulation of the evolutionary rate. Gene expression level is one of the strongest and consistent correlation between genomic data and evolutionary rate (Pearson correlation r of 0.53 and p-value less than 1e−9)4. Previous estimations have established that protein structural constraints could explain as much as 10% of the evolutionary rate in proteins5 but recent findings using mono and multiple-domain proteins indicate that structure-functional features and translation rates could have comparable contributions to explain evolutionary rates6. Most of these studies have been done describing the native state of a protein with a single structure. However, it is well established that native state of proteins are better represent by an ensemble of different conformers in dynamical equilibrium7. In this work we study how the presence of conformational diversity in mono and multi-domain proteins could influence the rate of evolution.

METHODS

To study this relationship we used a major update of the PCDB database (Protein Conformational Data Base)8. This database contains almost 8000 proteins with different degrees of structural diversity measured as the maximum root-mean-square deviation (RMSD) found between the different conformers for each protein. We have used a subset of the total number of proteins deposited in PCDB. This set consists in those proteins showing conformers with and without binding ligands. The RMSD was normalized for structural alignment length9. Each of these proteins was linked to Evola database10 of orthologs for human genes to estimate dN (ratio of the number of non-synonymous substitutions per non-synonymous site) as a measure of evolutionary rate using PAML 4 (Phylogenetic Analysis by Maximum Likelihood)11 with different codon models. We used CATH database12 for analysis of domains in a given protein. Domains was selected, for further dN estimation, using clustal omega13 protein alignments, and pal2nal14 facilities. Statistical analysis was done using R, package ppcor was used for partial correlation with expression level and boot for bootstrap analysis12,13,14,15. Microarray expression data from Affimetrix Human U133A was obtained from BioGPS.16

RESULTS & CONCLUSIONS

We found a negative correlation between dN (for orthologs between mouse and human) and the maximum RMSD for alpha carbons between protein conformers (Spearman rank correlation with a rho of -0.34 and a p-value less than 5 percent) for mono domain proteins in humans. This rho was tested with a bootstrap, the interval of confidence at the 95% level goes from -0.5 to -0.1. This correlation is independent of protein’s expression level (Partial Spearman rank correlation rho of -0.36 and P lesser than 5%) and doesn’t depend in the class C of CATH. However the level 4 of CATH, proteins with low content of secondary structure, don’t show this correlation between RMSD and dN.

This correlation become close to zero for proteins with multiple domains. It’s possible than homogenization of dN and RMSD occurs in proteins with more than two levels. Our results indicate that conformational diversity have an important role modulating protein evolutionary rates. We think that our findings could have important implications in the understanding of protein evolution process.

REFERENCES

AN EVOLUTIONARY LINK BETWEEN TWO APICOMPLEXAN PROTEIN FAMILIES WITH DISTINCT DISULFIDE BONDS

Dietlind L. Gerloff1, Edward L. Liu1, Eli Draizen1, Felicia D. Kemp2, D. Steven Hall1
Biomolecular Engineering Dept, University of California Santa Cruz1 and San Jose State University2.

The SRS surface proteins in Toxoplasma, and the "6-Cys" (PGSH) family in Plasmodia seem distantly related but their functions and disulfide bonds differ. To find clues about their evolutionary history, we screened 102,878 predicted proteins from Apicomplexan genomes sensitively with merged family HMMs. In the Coccidia (Toxoplasma/Neospora) we found newly evolved cysteine positions and a candidate "evolutionary link", with unusual resemblance to the Plasmodia/Piroplasmida homologs that may help us understand this phenomenon.

INTRODUCTION
Surface proteins in infectious pathogens evolve rapidly and different species feature remarkably different arsenal of them. Fold similarity can be difficult to detect by sequence-based methods, in this category, which contains many crucial targets for vaccine development. In 20051 we detected distant homology of the malarial transmission-blocking antigen Pf48/45 to an experimentally better tractable family of surface proteins in a different Apicomplexan species: the SRS (SAG1-related sequences) family in Toxoplasma gondii2. To make the sub-significant similarity signal at the time more apparent, and to model the structure of the characteristic double-domain repeat in the targeted malarial "6-Cys" protein family (which we prefer to call PGSH: Plasmodium gametocyte surface homology proteins), we excised exceptionally variable insertions. Our confidence in the model increased further when we realized that the SRS fold-based prediction would effortlessly support the distinct disulfide bonds asserted in the PGSH proteins1. The phenomenon is well known (e.g. from immunoglobulin structures) but it is often underappreciated, and it results in family-specific conserved cysteines that are misinterpreted to be essential. But how do disulfide bonds "change position" during evolution, and is there any reason (besides academic curiosity) why we should care? In this case, the best known representatives of the SRS and PGSH families have very different functional roles1,2. Thus we embarked on the sequence-based pursuit of their most ancestral-like forms primarily motivated by the hope to discover a functionally interesting protein.

METHODS
After revalidating the fold prediction in light of new sequences and structures (not shown), we undertook a large-scale hidden Markov Model (HMM) screen of the 102,878 putative proteins in EuPathDB, from completely sequenced Apicomplexan genomes (Nov 2011). We used a HMMER3.03 (E-value(best) ≤ 0.1) SAMv3.54 (E-value(match) ≤ 1.0) method combination pipeline for maximum sensitivity and reliability. After filtering at ≤ 95% sequence identity we identified a 74 position core segment in the piled-up matches, and used this for phylogenetic clustering (with TreeTop).

RESULTS & CONCLUSIONS
Among other interesting findings, our project identified an atypical SRS/PGSH homolog in the Tissue Cyst forming Coccidia (which include T. gondii and N. caninum) that may resemble the "evolutionary link" we sought. The core segments of the ortholog pair (found with E-values around 10^−2−10^−7 by multiple methods) cluster away from the other Coccidian (SRS) matches (Figure 1A). Also the only distance-plausible disulfide bond in the N-terminal (non-core) region is PGSH-like (Figure 1B) and profile-based methods favor PGSH over SRS. While functional studies are pending, this and molecular modelling help explain how disulfide bonds may "move" (and why we care).

REFERENCES
SIZE MATTERS: SURFACE HYDROPHOBICITY INDEX (SHI) DESCRIBES THE IMPACT OF THE SIZE OF INTERFACE AREA ON OLGOMERIZATION DRIVEN BY HYDROPHOBIC EFFECT

Izabella A.P. Neschich1,2, Fábio de Moraes1,2, José A. Salim1, Ivan Mazoni1, José G. Jardine1, and Goran Neschich1,3

National Agricultural Information Technology Research Center (EMBRAPA-CNPTIA), Biology Institute at State University of Campinas (UNICAMP)1. neschich@cnptia.embrapa.br

We present an extensive study of protein-protein interfaces in a non-redundant dataset (DS95) containing 6.931 PDB chains which highlights the relationship between hydrophobic effect and oligomerization of proteins by using a newly defined Surface Hydrophobicity Index (SHI). We found that 92% of all DS95 chains have their interface more hydrophobic than the rest of a solvent-exposed surface. For interfaces larger than 3,000 Å², this percentage goes to 98%. Additionally, positive correlation for DS95 ΔSHI values (representing the interface hydrophobicity) and size of the interface area (Å²) was detected.

INTRODUCTION
The details of how proteins are interacting between themselves are not fully understood, but many hypotheses have been proposed. Traditional studies have shown that the hydrophobic effect plays a significant role in oligomerization [1,2]. Argos and coworkers [3] described that the largest hydrophobic patches at surface of most complexes analyzed were coincident to interface regions. We decided to design a new approach that would define how to measure the hydrophobicity of interfaces and be capable of assessing how important and wide spread is such contribution for assembly of complexes in a known protein structure universe. To achieve this, we have defined a curated dataset of non-redundant protein chains from PDB (DS95, 6,931 chains) and a specific parameter related to hydrophobicity: The Surface Hydrophobicity Index (SHI). The principle considered here was that if the hydrophobic effect is a driving force for protein oligomerization, most of DS95 interface areas should be slightly more hydrophobic than the remaining surface.

METHODS
We created four distinct datasets using a set of filters to choose non-redundant protein-protein complexes from PDB: chains > 50 amino acids, interface area > 200 Å²; no fragment proteins; X-ray only with resolution < 3.0 Å; no DNA/RNA. The sequence identity in datasets were DS100(%), DS95(%), DS70(%) and DS30(%). There were no statistically significant differences regarding results obtained among different datasets. Both DS100 and DS95 (9.009 and 6.931 chains, respectively) are the largest protein-protein interaction datasets described in the literature so far. We introduced a parameter, derived from three known scales (Kyte-Doolittle, Eisenberg and Engelman), normalizing each amino acid residue Hydrophathy by their effective accessible surface area. Also, a new parameter was created, reported in a per chain fashion: Surface Hydrophobicity Index (SHI) which describes the cumulative surface Hydrophobicity for a selected chain in two flavors: isolated chain and chain in complex with another chain. Low SHI values are indicators of hydrophobic protein surfaces and vice versa. The ΔSHI (SHI_complex – SHI_Isolated) represents the interface hydrophobicity and a positive ΔSHI indicates that the interface area is more hydrophobic than the remaining protein surface area. We decided to design a new approach that would define how to measure the hydrophobicity of interfaces and be capable of assessing how important and wide spread is such contribution for assembly of complexes in a known protein structure universe.

RESULTS & CONCLUSIONS
A positive correlation was found to exist among ΔSHI value and the ratio between corresponding interface size and the total surface area size (both for protein chain and complete protein/oligomeric complexes). This implies that as the size of the interface grows, so it does the area of hydrophobic residues that compose selected interface which needs to be buried during complex formation. Membrane proteins have no correlation among ΔSHI and interface to surface area ratio. Slightly more than 92% of all studied interfaces obey the rule: ΔSHI > 0, and for interfaces of the most frequent size (>3000Å²) in the DS95 this percentage goes to more than 98% (Figure 1).

The cases which do not obey ΔSHI>0 rule were frequently found to be proteins having significantly smaller than the average interface sizes or membrane proteins. More importantly, a total of 99.9% of complexes where core residues are found to make part of the interface (85% of the DS95 complexes), obey either one of the three following rules: dSHIp=0 or dSHI<0, suggesting the high degree of occurrence of cases where hydrophobic effect has a role in protein complex formation (dSHIp corresponds to SHI value calculated for protein conglomeration considered complete and dSHIcore corresponds to the SHI value where the interface is identified with the region where amino acids have completely lost access to a solvent).

We believe that the entropy change in the system during water expulsion and hydrophobic effect is the driving force that leads one chain to approximate another chain in a specific region.

REFERENCES
EVOLUTION OF FUNCTIONS IN DOMAIN STRUCTURES–
EXPLORING SHIFTS IN FUNCTIONAL SITES

Benoit Dessailly, Alison Cuff, Natalie Dawson, Robert Rentzsch, Christine Orengo
Department of Structural and Molecular Biology, University College London, UK, c.orengo@ucl.ac.uk

INTRODUCTION: Analyses of domain structure superfamilies in the CATH-Gene3D classification have revealed functional shifts amongst domain relatives in some of the most highly populated and structurally diverse superfamilies (1). Functions can only really be attributed to the whole protein, but since there are many fewer domain families (of the order of thousands) than protein families (likely to be tens of millions) it would be helpful to understand the functional roles of domains and how these change in different protein contexts. This is particularly important because the international genome sequencing initiatives are yielding millions of protein sequences with no experimental characterisation.

Most proteins comprise domains from families already characterised in CATH-Gene3D, SCOP, Pfam and other resources. However, these primary domain building blocks are often combined in different ways. We are striving to develop a functional ‘grammar’ of domain families as an important step towards determining the functions of the complete proteins. To do this we need to determine whether relatives have a fixed functional role or can adapt new functions.

Recent sub-classification of CATH superfamilies into functional families (FunFams) using sensitive HMM based protocols (2) has allowed us to group relatives sharing very similar functions. In some cases it is clear that the functional roles of relatives are unchanged despite changes in multi-domain context. These domains are found in many proteins because they have generic roles which can easily be exploited in different contexts. In other superfamilies, changes in the functional role of the domain itself are observed, mediated by structural variations or shifts in residues.

We have performed extensive analyses to determine how functional sites change across superfamilies. We identify a surprising number of functional residue shifts in enzyme superfamilies. Analysis of conserved interface patches between different functional families have also shown that whilst some domain superfamilies preserve a common site, others exploit multiple sites across a large area on the domain surface.

METHODS: We have used a modified HMM based version of the GeMMA agglomerative clustering algorithm (2) to subclassify domain superfamilies in CATH-Gene3D into functional families (FunFams). Multiple alignments were generated in each superfamily using MAFFT (3) and scorecons (4) was used to identify highly conserved functional positions. Validation of the purity of the FunFams came from observing that the conserved sites correlate well with known catalytic residues from CSA or interface sites from IBIS. A suite of programs (SiteMapper, SiteShift) compare sites across functional families and identify shifts in functional residues and protein interface patches.

RESULTS & CONCLUSIONS

We have detected examples of extreme shifts in the nature and location of conserved residues in enzyme active sites in different functional families of some large enzyme superfamilies (see figure 1 for an example). Our strategy has been endorsed by validating against the functionally diverse TPP superfamily, a well studied family frequently used in protein design to engineer new chemistries and specificities. We will present summary data on the functional site changes, together with selected examples and characterisation of changes observed.

We also observe changes in the locations of protein interfaces. In some superfamilies practically the entire protein surface is exploited in different interactions. Whilst in other superfamilies, one or a few sites are highly conserved. This restriction provides a means for regulating protein interactions as it obviously excludes binding by multiple partners at the same time. Therefore, structural data on interfaces can provide valuable dynamic information on protein networks. We will describe how we plan to use this structural information to annotate protein networks in CATH-Gene3D.

REFERENCES

AN INTERACTIVE ONLINE PLATFORM FOR STRUCTURE MODELING OF G-PROTEIN COUPLED RECEPTORS

Dorota Latek1,2*, Paweł Pasznik1, Stawomir Filipěk1
Faculty of Chemistry, University of Warsaw1, International Institute of Molecular and Cell Biology in Warsaw2.
*dlatek@iimcb.gov.pl

We present an interactive platform for modeling structure of GPCRs and its complexes. Beyond a fully automatic processing we enabled interactive change of various settings: sequence profile distribution, loop-spanning residues and the modeling method. We combined wide-employed methods, such as BLAST, Muscle, Rosetta, Modeller together with the membrane-related scoring functions and a docking procedure.

INTRODUCTION
G-protein coupled receptors are targets of nearly half of drugs at the current pharmaceutical market. Despite the great interest only few structures are known to date, because of experimental difficulties in finding the best environment, mutations and ligands to stabilize these transmembrane proteins in crystals. In contrast to time-consuming experimental methods, computational biology offers fast and accurate methods for homology modeling of GPCRs which structure consists of a common seven transmembrane helices fold. The main area of research associated with GPCRs is drug discovery requiring a great amount of computational resources. Here, we propose an integrated online platform for modeling of GPCRs and their small-molecules complexes. The platform will significantly decrease the time of structural analysis and provide computational resources needed in large scale biological projects.

METHODS
The platform1 (Figure 1) employs sequence profiles generation2, anchored aligning of sequence profiles3, assessment of template-target alignment, an extensive loop modeling performed by two different methods (Rosetta4 and Modeller5) and final model assessment based on statistical potentials6. The platform can be run in two modes: fully automatic and interactive one in which a user can change settings, modeling methods as well as the template-target alignment and the templates subset. The final protein model is refined in the all-atom force field with hydrogen atoms included and thus can easily serve as a receptor in a fully flexible docking performed online. Small molecules used in the docking can be either agonist or inverse agonists since the activation state of GPCR is taken into account during the model building procedure. Final results of the modeling can be examined online by a Jmol applet. The platform was built on a Django framework combined with python applications supported by a Biopython7 library.

RESULTS & CONCLUSIONS
The platform was successfully tested in the last GPCR Dock 2010 competition in modeling of the Dopamine D3 receptor in complex with a small molecule eticlopride8. We provided the molecular biology community a valuable computational tool accessible online. Due to efficient combination of various modeling techniques we omitted the time-consuming molecular dynamics step usually used in studies of transmembrane proteins. The platform can be used not only by computational biologist but also by experimentalists to verify their data.

REFERENCES

Figure 1. GPCR modeling platform – screenshot of the main form.
FORMAT DÉJÀ VU: PDBx/mmCIF, THE NEW DATA FORMAT
FOR THE WWPDB

John Westbrook1*, Helen M. Berman1, Jasmine Young1 & Gerard J. Kleywegt2
RCSB PDB, Rutgers University, Piscataway, NJ USA1 and PDBe, EMBL-EBI, Hinxton, Cambridge, UK2.
jwest@rcsb.rutgers.edu

At a September 2011 meeting held at PDBe, the wwPDB met with key software developers to discuss the limitations of the PDB file format in its ability to support data for large structures, complex chemistry, and new and hybrid experimental methods. Rather than designing an entirely new format, the consensus among developers was to actively switch to the PDBx/mmCIF format, particularly for use in deposition. A meeting outcome was the formation of this working group for the adoption of the PDBx/mmCIF format for deposition and as an exchange format between programs for macromolecular crystallography. Making it easier to capture more information in deposition and archiving will be part of this process.

The PDB archive is an international repository managed by the Worldwide Protein Data Bank (wwPDB) [1], wwPDB members curate, annotate, and distribute PDB data, while making great efforts to maintain consistency and accuracy across the archive. As the PDB grows, new structures and new technologies impose new demands on how structures are represented. To address these demands, the wwPDB will adopt the mmCIF/PDBx [2,3] format to replace the current PDB format.

The wwPDB has provided data in PDBx/mmCIF format for many years; however, in order to maintain reversible format translation with the PDB format, much of the expressivity of the PDBx/mmCIF has not been exploited. In choosing PDBx/mmCIF as the primary format for both data deposition and data delivery, PDB data will be free of the many format restrictions imposed by the older record-oriented format. The wwPDB will take advantage of the flexibility provided by the PDBx/mmCIF framework to represent larger molecular systems and to capture ever-evolving methodological details. This format flexibility will also be used to produce a data presentation style that preserves the simple and popular organization of the current PDB format.

In this presentation, we discuss the new features and content of the PDBx/mmCIF format and software tools that have been developed to use mmCIF and the supporting PDB Exchange Data Dictionary. We will also report on the progress of the Working Group for PDB Deposition using PDBx/mmCIF. This group includes developers from several major crystallographic and NMR structure determination program systems.

The wwPDB members are: RCSB PDB (supported by NSF, NIGMS, DOE, NLM, NCI, NINDS and NIDDK), PDBe (EMBL-EBI, Wellcome Trust, BBSRC, NIGMS, and EU), PDBj (NBDC-JST) and BMRB (NLM).

REFERENCES
HOW DOES KCNE1 REGULATE THE KV7.1 POTASSIUM CHANNEL? MODEL STRUCTURE, MUTATIONS AND DYNAMICS OF THE KV7.1-KCNE1 COMPLEX

Yana Gofman†, Simona Shats§, Bernard Attali†, Turkan Haliloglu‡, and Nir Ben-Tal†‡*
† George S. Wise Faculty of Life Sciences, and § Sackler Faculty of Medicine, Tel-Aviv University, Israel. ‡ Bogazici University, Bebek-Istanbul, Turkey. *NirB@tauex.tau.ac.il

The voltage-gated potassium channel Kv7.1 and its auxiliary subunit KCNE1 are expressed in the heart and give rise to the major repolarization current. The interaction of Kv7.1 with the single transmembrane helix of KCNE1 considerably slows channel activation and deactivation, raises single-channel conductance and prevents slow voltage-dependent inactivation. We built a Kv7.1-KCNE1 model-structure. The model-structure agrees with previous disulfide mapping studies and suggests molecular interpretations of electrophysiological recordings of two KCNE1 mutations that are presented here. We conducted elastic network analysis of Kv7.1 fluctuations in the presence and absence of KCNE1. The analysis suggests a mechanistic perspective on the known effects of KCNE1: slow deactivation is attributed to the low mobility of the voltage sensor domains upon KCNE1 binding; abolishment of voltage-dependent inactivation could result from decreased fluctuations in the external vestibule; amalgamation of the fluctuations in the pore region is associated with enhanced ion conductivity.

INTRODUCTION
Despite the biological significance of the Kv7.1-KCNE1 interaction, its exact nature has yet to be elucidated. Kang and co-authors predicted the structure of the Kv7.1-KCNE1 complex1,2, and our preliminary analysis has shown that their model is incompatible with evolutionary information in that the most conserved amino acids of KCNE1 are exposed to the lipid and the variable residues are buried at the interface with Kv7.1, in contrast to the typical conservation pattern. Therefore, herein we attempt to provide further insight into the Kv7.1-KCNE1 interaction, using the available model-structure of Kv7.12 and various computational protein-protein docking techniques in combination with evolutionary conservation data. The model structure is consistent with available disulfide links data. To test it further we examined the electrophysiological properties of Kv7.1 co-expressed with two KCNE1 mutants; overall, the experimental findings are consistent with our model-structure of the complex. Analysis of the model-structure, using elastic network models, suggested a molecular basis for alterations in the dynamics of the Kv7.1-KCNE1 complex in comparison to the isolated Kv7.1 channel, and interpretation of clinical mutations.

METHODS
Model building. The 3D structure of the transmembrane segment of KCNE1 was modeled using the Rosetta Membrane ab-initio protocol3. Various model structures of this segment were docked to the available homology model of Kv7.12 using PatchDock4. The model-structures were evaluated based their membrane topology and hydrophobicity and evolutionary conservation profiles. The anticipation was that polar and conserved residues would be buried at the protein-protein interface and hydrophobic and variable residues would face the lipid. One model was selected. Elastic network models. The model-structure of the Kv7.1-KCNE1 complex was analyzed using the Gaussian network model (GNM)5 and anisotropic network model (ANM)6.

RESULTS & CONCLUSIONS
We present elastic network analysis of the equilibrium dynamics of our model-structure of the Kv7.1-KCNE1 complex. Four main modes of motion were detected. The hinge residues of one of them are presented in Figure 1 to demonstrate the usefulness of the analysis to suggest functional interpretations of mutations. Overall, the model-structure agrees with the evolutionary conservation pattern and available disulfide mapping studies. Furthermore, it suggests molecular interpretation for our electrophysiology studies of the channel co-expressed with two KCNE1 mutants (e.g., Fig. 1). The elastic network analysis of the model-structure revealed interesting notions on channel deactivation, inactivation and gating, proposing important roles for D317 and E295 in channel inactivation and for G272, V310 and T311 in conductivity. Moreover, it proposed a mechanistic perspective of how KCNE1 binding affects channel dynamics. Slow deactivation was attributed to the low mobility of the voltage sensor domains upon KCNE1 binding; abolishment of voltage-dependent inactivation resulted from the decreased fluctuations in the external vestibule; amalgamation of the fluctuations in the pore region was associated with the enhanced conductivity of the channel.

REFERENCES
Loops are irregular structures which connect two secondary structure elements in proteins. Here we examine the span of a loop, the distance between its two end points. We find that the distribution of loop span appears to be independent of the number of residues in the loop, and is also unaffected by the secondary structures at the end points. We demonstrate that the span distribution can be described by a random fluctuation model based on the Maxwell-Boltzmann distribution. We also show how normalised span (loop stretch) is related to the structural complexity of loops: Highly contracted loops are more difficult to predict than stretched loops.

**INTRODUCTION**

Protein loops are patternless secondary structure elements. They are generally located in solvent exposed areas and often play important roles. Despite the lack of patterns, loops are not totally random. Here, we focus on a specific local property of protein loop structure: the distance between the two terminal Cα atoms of the loop, which we refer to as “span”. We demonstrate that the observed span distribution can be largely explained by a simple model based on the Maxwell-Boltzmann distribution. It is widely believed that the accuracy of loop structure prediction depends on the number of residues, i.e. the larger the number of residues, the more difficult to predict.

Here, we introduce the normalised span which indicates how stretched a loop is. By examining an ab initio loop structure prediction method, we show that loops of their maximum spans are predicted accurately regardless of the number of residues.

**METHODS**

A loop structure is defined as any region that is neither helix nor strand.

Loop span (l) refers to the physical spatial distance between two end Cα atoms of a loop. Maximum loop span (l_max) is furthest that a set of residues can spread.

\[
l_{\text{max}}(n) = \begin{cases} \\
\gamma \cdot (n/2 - 1) + \delta \text{ if } n \text{ is even} \\
\gamma \cdot (n - 1)/2 \text{ if } n \text{ is odd} 
\end{cases}
\]

where \( n \) is the number of residues, \( \gamma = 6.046 \text{Å} \) and \( \delta = 3.46 \text{Å} \). Loop stretch (\( \lambda \)) is the normalised loop span, which is the observed loop span between two Cα atoms at each end of a loop divided by the loop’s maximum span.

\[
\lambda = \frac{l}{l_{\text{max}}}
\]

**RESULTS & CONCLUSIONS**

The nature of the span distribution is broadly similar across different secondary structures which a loop connects (anchors). In particular, the most highly frequent span appears to stay the same (approx. 14Å) irrespective of the anchor type (Figure 1). These observations suggest that the loop span is distributed independently of local anchor structures. We also split loop spans in terms of the number of residues (Data not shown here). Despite the range of the number of residues, the mode still stays the same across loops containing varying numbers of residues.

We hypothesised that the two end points of loops freely fluctuate in three dimensional space, following the Maxwell-Boltzmann distribution. As shown in Figure 1, the simulated span distribution fits the span distributions.

We then examined the effect of loop span on loop structure prediction. We calculated the normalised loop span (loop stretch), and tested MODELLER on contracted and fully stretched loops. We found that fully stretched loops are accurately predicted regardless of the number of residues (Figure 2).

**FIGURE 1.** The loop span distributions in terms of the anchor secondary structures do not show differences except for loops linking anti-parallel beta strands. The Maxwell-Boltzmann distribution largely explains the loop span distribution.

**FIGURE 2.** Accuracy of protein loop structure prediction methods does not only depend on the number of residues, but also on loop stretch.

It has been believed that the accuracy of protein loop structure prediction is related to the number of residues. We show that the primary determinant of protein loop structure prediction accuracy in fact appears to be loop stretch.

**REFERENCES**

**Introduction**

The Herpes viruses are important human pathogens that can cause mild to severe lifelong infections with high morbidity. Glycoproteins gB and gH-gL are highly conserved cell entry machinery which are involved in attachment and fusion of herpes virus to the host cell. gB is a homodimer/trimer with structural characteristics to undergo conformational rearrangement when triggered, thus inferred to be the effector of viral fusion. Whereas gH-gL glycoproteins are the heterodimer complex proposed to be the activators of gB glycoprotein possibly through direct binding. Critical dependence of the herpes virus on the formation of this gB-gH-gL complex, non-formation of which inhibits the viral entry to the host cell making this interface a promising anti herpes drug target. In this present study we reported the key amino acid residues involved in this complex formation along with elucidation of possible mode of action for the most promising anti-herpes drugs.

**Methods**

The key amino acid residues in the gB-gH-gL complex formation were identified based on Protein-protein docking interactions study conducted using Hex 6.3 standalone software. Structure based coupled with ligand based virtual screening was conducted to identify the potential inhibitors for gB and gH-gL complex using Arguslab 4.0.1 and ZINC database respectively. Flexible and semi-flexible docking studies were conducted for the most promising compounds targeting the key amino acid residues to elucidate their capability to attenuate this complex formation using Autodock 4.2. Hydrogen bonding network along with other interactions was studied using Discovery studio Visualizer 3.0.

Pharmacological features predictions for the promising compounds was carried out based on descriptors values found in ZINC database and LAZAR online server was utilized for prediction of toxicity of the compounds.

**Results & Conclusions**

We have identified Val 342, Glu 347, Ser 349, Tyr 355, Ser 388, Asn 395, His 398 and Ala 387 amino acid residues of gH-gL complex and Lys 435, Arg 418, Gin 438, Pro 439, Leu 399, Glu 401 and Gly 437 of gB protein (Figure 1) are found to be active during their active complex formation. Firstly, we conducted a semi-flexible docking study targeting these key residues for the published ethno compounds which showed significant in-vitro & in-vivo inhibition against Herpes virus as reported elsewhere.

(3-chlorophenyl) methyl 3,4,5-trihydroxy benzoate (CPMTHB) was found to be the most promising among the published compounds based on its free binding energy and pharmacological features. This compound was further used for ligand based virtual screening to identify much better compounds.

All the 505 compounds found through ligand based virtual screening studies were showed significantly good binding affinity towards gB and gH-gL active residues in our structure based virtual screening, suggesting their capability to inhibit viral entry into the host cell targeting this conserved cell entry machinery.

**References**

PREDICTING KINKS IN TRANS-MEMBRANE HELICES

Henry Wilman1,2*, Dr. Charlotte Deane1,2 & Dr. Jiye Shi1
Department of Statistics and Systems Approaches to Biomedical Science Industrial Doctorate Centre, Oxford University, UCB CellTech1, henry.wilman@hertford.ox.ac.uk

Alpha-helical bundles make up the majority of membrane proteins, and are both medicinally relevant and biologically interesting. Given the few experimentally determined structures available, structure prediction is an important tool in these areas. We have focused on the phenomenon of kinks in trans-membrane helices.

An automated kink detector has been developed. The information extracted by this from a non-homologous set of membrane protein structures has been used to investigate the nature of kinks. This knowledge has been used to inform the design of a program to predict the position and size of kinks in trans-membrane helices.

RESULTS & CONCLUSIONS
Analysis of the sequences around these kinks identified Proline as a major cause of kinks, as seen in previous studies [2-5]. We also saw a number of other amino acid effects, including some amino acids preferring to be on the compressed or stretched side of the kink. In addition, by utilizing iMembrane, a homology based tool for positioning membrane protein structure in the membrane, we saw that kinks occurred more frequently in certain parts of the membrane.

REFERENCES

INTRODUCTION
Here we present a knowledge-based predictor of kink position and size, using input of amino acid sequence and membrane position information derived from iMembrane [1].

Alpha-helical bundles make up a majority of integral membrane proteins, and include G-protein coupled receptors and ion channels. These are both medicinally relevant.

The difficulty of experimentally determining their structure makes modelling these membrane proteins an important tool in drug design, as well as a tool to probe their structural biology.

Trans-membrane helices are often kinked, and these kinks have been shown to be functionally important. Sequence mutations have been associated with diseases, and molecular dynamics simulations have demonstrated relationships between amino acids and function [6,7].

METHODS
Here we demonstrate a knowledge based tool to predict the size and position of kinks in trans-membrane helices using sequence information.

Initially we studied a non-homologous set of helices from the PDB, using our own method to automatically identify kinks, identifying a residue on the stretched side of the helix as the kink point.

Using relationships observed in this analysis, we constructed a score based predictor of kink position and size, using input of amino acid sequence and membrane position information derived from iMembrane.

[Image of Helices in the membrane. Axes are fitted to segments of the helix, and angles are calculated between consecutive segments. The largest angles are identified as kinks.]
**FACTORS AFFECTING MEMBRANE PROTEIN THREADING**

Jamie Hill¹ & Charlotte M. Deane¹,*

Dept. of Statistics¹, University of Oxford. *deane@stats.ox.ac.uk

Membrane proteins comprise up to 50% of drug targets, but only ~2% of structures in the pdb. In template-based modelling of membrane proteins, the primary determinant of model quality is the threading between the target and template. We determine the factors that affect threading quality in the membrane regions, and incorporate this knowledge into a threading method. Despite its simplicity, on a test-set of 190 membrane protein pairs, our method outperforms PROMALS, MSAprobs, mafft, muscle, clustalΩ, PSI-Coffee and hhsearch.

**INTRODUCTION**

Membrane proteins represent approximately 30% of the human proteome, yet it is estimated that there are only ~350 unique membrane protein crystal structures. For comparison, there are presently >45 000 unique crystal structures in the Protein Data Bank. In the absence of crystal structures, homology models for a sequence may be built from the structures of related proteins. The quality of such models depends primarily on the alignment that maps the sequence of interest onto the template structure.

Previously, we constructed a pairwise threading method that used a different substitution table for each combination of secondary structure and membrane positioning in a protein¹. Here we use this pairwise method as a foundation for a membrane multiple sequence alignment threader.

We demonstrate that a multiple-sequence approach involving membrane specific information can align membrane proteins better than the best conventional approaches, and discuss how homolog-selection, phylogeny inference and secondary structure propensities affect threading performance.

**METHODS**

We align homologous sequences for our target and template using a progressive alignment strategy. Our pairwise alignment method selects a different substitution table for each residue based on a structurally annotated sequence of known structure. However, many of the homologs to be aligned have no known structure. To compensate for this, we first align every homologous sequence in a pairwise-manner to the structure, and transfer structural annotation between aligned pairs of residues. By this technique we can effectively perform a sequence-structure alignment at each stage of the multiple-sequence alignment.

Our multiple sequence aligner optimizes the T-Coffee consistency criterion³, using a sequence-library composed from the above pairwise sequence-structure alignments.

**RESULTS & CONCLUSIONS**

On a test set of 190 target-template pairs our method outperforms other multiple-sequence methods including MSAprobs and clustalΩ, as well as threading methods such as PROMALS and hhsearch.

```latex
\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Two views of a model (orange) of an 8-stranded outer-membrane protein (green). The template structure used to build the model shares 9% sequence identity with the target.}
\end{figure}
```

Models of transmembrane domains built from our alignments have the lowest RMSD and highest GDT_TS scores of any methods tested. Figure 1 shows two sides of a model (orange) built for an 8-stranded outer membrane protein. The target and template have less than 10% sequence identity, but the model has a GDT_TS of 0.67 over the aligned residues.

**REFERENCES**

MULTI-CONFORMER CONTACT NETWORK ANALYSIS PROVIDES THE STRUCTURAL AND FUNCTIONAL BASIS FOR NMR RELAXATION EXPERIMENTS

Henry van den Bedem1,2, Gira Bhabha2, Kun Yang1, Peter Wright1 & James Fraser5

Joint Center for Structural Genomics, Stanford Synchrotron Radiation Lightsource 1
Dept of Cellular & Molecular Pharmacology, UCSF 2
Institute for Computational and Mathematical Engineering, Stanford University 1
Dept of Molecular Biology, The Scripps Research Institute 6
vbedem@stanford.edu; james.fraser@ucsf.edu

We developed a multi-conformer Contact Network Analysis (mcCNA) algorithm to identify connected pathways of interacting residues in X-ray crystallography data. Our method is based on propagation of overlapping van der Waals radii through residues with alternative conformations. We identified networks implicated in catalysis for Escherichia coli dihydrofolate reductase (ecDHFR) and cyclophilin A (CypA) in X-ray data collected at ambient, but not cryogenic, temperatures that provide a structural basis for NMR relaxation data and explain how mutations affect both catalytic rate and protein conformational dynamics. Thus, advanced algorithms to interpret electron density can foster novel, verifiable hypotheses around functional mechanisms with broad implications for protein design and drug discovery.

INTRODUCTION

Proteins populate fluctuating conformational ensembles which mediate biological functions. While Nuclear Magnetic Resonance (NMR) spectroscopy experiments yield valuable insight into the picosecond to seconds time-scale dynamics of the ensemble, obtaining a structural interpretation from these experiments remains a challenge.1,2 Relaxation dispersion experiments determine populations and rates of interconversion of excited states on the micro-milli-second time-scale, but only provide indirect information of structural changes by chemical shift differences between the ground and excited states. Developing a structural ensemble representative of the data is impeded by the computational cost of sampling a prohibitively high-dimensional space and the indirect relationship between chemical shifts and structure.3 Interpreting chemical shift changes from mutations to obtain insight in long range interactions is difficult, and results are often confounded by unrelated fluctuations. These limitations can be overcome by interpreting NMR results with the aid of NMR relaxation data and explaining visually identifying correlated motions relate to function. Thus, advanced

METHODS

We obtain multi-conformer protein crystal structures from our previously developed robotics-inspired algorithm qFit.4 This procedure computes an optimal model from an electron density map taking fully into account main- and side-chain heterogeneity. The mcCNA algorithm treats residues as ‘falling dominoes’: a pathway propagates between a pair of residues through overlapping van der Waals radii whose clashes can be relieved by selecting an alternate conformation. By tracking such ‘clash-and-relieve’ interactions, pathways of interacting residues are identified.

RESULTS & CONCLUSIONS

With mcCNA we find that 1) Cryogenic conditions greatly reduce the number and lengths of pathways despite a higher packing density5 compared to ambient temperature data. 2) Dynamic networks in ambient temperature X-ray data provide a structural characterization of ensemble dynamics consistent with NMR relaxation-dispersion and chemical shift data in the enzymes cypA and ecDHFR.

FIGURE 1. Coupling between the F-G loop and residues 63-65 in the adenosine binding domain of a (E:NADP+:FOL) complex. NADP+ on left, Folate on right in molecules. (a) Residues obtained from pathways visiting the F-G loop are colored in red. The thickness represents how often the residue occurs, (b-c) Chemical shift differences from WT and a G121W mutant are shown in red with (b) and without (c) NADP+. Note the absence of chemical shifts in the adenosine domain (top of the molecule) in (c). The thickness visualizes deviation from the average chemical shift difference. The mutation site is depicted with an arrow.

Analyzing 30 matched pairs of data sets collected at cryogenic and ambient temperatures we found that cryogenic data sets on average yield 56% fewer pathways that are 41% shorter than their ambient temperature counterparts. X-ray data from a (ecDHFR:NADP+:FOL) complex confirms that the F-G loop is coupled to the adenosine binding domain (Figure 1a,b), and furthermore establishes that pathways run exclusively through the cofactor, consistent with NMR data (Figure 1c).

Ambient temperature X-ray data supplements cryogenic data by enriched heterogenous features. Accordingly, analysis of a multi-conformer model from a single X-ray experiment offers a new strategy, complementary to long-timescale molecular dynamics, Rosetta sampling and comparison of multiple independent crystal structures, to derive a structural basis for conformational dynamics of proteins and to develop hypotheses about how these motions relate to function.

REFERENCES

THREADING MULTIPLE SUBUNITS INTO AN EM DENSITY MAP: APPLICATION TO THE 26S PROTEASOME

Keren Lasker 1,2,*, Daniel Russel 1, Jeremy Phillips 1, Friedrich Förster 1, Wolfgang Baumeister 3, Haim Wolfson 2, & Andrej Sali 4

University of California San Francisco, 1 Tel Aviv University, 2 Stanford University (current address), 3 Max Planck Institute of Biochemistry 1, keren@salilab.org

A limiting factor in structural interpretation of an intermediate resolution cryo-EM density map of an assembly is the structural coverage of its subunits. When the structural coverage is not complete, investigators will often pursue a higher resolution density map. However, intermediate resolution density maps along with their sequences contain a significant amount of information about the subunit native structures. Here, we present MultiThread, a method for simultaneously threading multiple protein subunits onto a discrete representation of their assembly density map, while satisfying restraints derived from the density map, the subunit sequences, and proteomics experiments. The method was successfully applied to solve the molecular architecture of the 18 subunits comprising the regulatory particle of the 26S proteasome.

INTRODUCTION

Electron cryo-microscopy (cryo-EM) has become a standard method for structure determination of macromolecular assemblies. Recent technological advancements have made it common to determine intermediate resolution maps (6-10Å), in which structural feature such as domain boundaries and secondary structure positions can be apparent. A limiting factor in structural interpretation of such maps is the structural coverage of its subunits. Many complexes determined at intermediate resolution are missing complete structural coverage, such as the proteasome, the ribosome, the skeletal muscle calcium release channel, and numerous viruses.

A common approach for interpretation of a density map is the fitting of its subunits into the density map in a consistent fashion (ie., avoiding clashes between the subunits and consistently with other experimental data). This approach is limited to cases for which atomic structures are available for all subunits. Integrative structure determination approaches have proven useful for providing a comprehensive structural picture of a protein assembly that cannot be derived from any single method alone 1-3. Specifically, imposing restraints derived from the density map and protein sequences of individual subunits, in addition to sparse experimental data derived from residue cross-links and proteomics, can result in a complete interpretation of the density map at the resolution of a single residue.

Given the ambiguity and incompleteness of the data sets listed above, many possible structures may satisfy the data. Generating a complete ensemble of structural models using traditional sampling techniques is difficult due to the high dimensionality of the search space and the large number of local minima. Here, we describe a divide-and-conquer sampling technique that can enumerate all possible solutions at a predefined resolution, given an intermediate resolution density map. This approach was recently used to determine the complete molecular architecture of the 26S proteasome 4.

METHODS

We have developed MultiThread, a method for predicting the architecture of protein subunits and their assembly, given their amino acid sequences, an EM map of their assembly, a database of known template structures, and sparse proteomics data. MultiThread hierarchically “threads” the subunit sequences through density anchor points using a data-dependent scoring function to estimate the quality of the matching. The objective is to find all threading solutions that agree with the data sufficiently well.

To demonstrate the performance of our approach, we generated a benchmark of 10 macromolecular complexes, tested under multiple combinations of structural coverage and sparse data. The results suggest that the molecular architecture can be determined at single residue resolution even when some individual subunits are not represented at atomic resolution, given sufficient data from proteomics and a density map at better than 10 Å resolution.

RESULTS & CONCLUSIONS

In eukaryotes, the ubiquitin–proteasome pathway regulates fundamental cellular processes such as DNA repair, signal transduction, and protein quality control. MultiThread was successfully applied to solve the molecular architecture of the 18 subunits comprising the regulatory particle of the 26S proteasome, which resides at the downstream end of the pathway and executes chemical reactions that degrade defective proteins. X-ray structures, comparative models, and coarse-grained representations of the constituent domains were assembled based on a cryo-electron microscopy map at 8.4 Å resolution, difference maps of 2 single-protein knockouts, 15 residue-specific chemical cross-links, and a large number of proteomics experiments (Figure 1). The resulting structure is informative about the molecular mechanism of the 26S proteasome.

ACKNOWLEDGMENTS

NIH Grants R01 GM54762, U54 RR022220, and R01 GM083960.

REFERENCES


FIGURE 1. Cluster analysis and structural model of the 26S proteasome (adapted from 4).
**INTRODUCTION**

In the past few years it has been established that a process of histone demethylation by recently discovered histone demethylation enzymes inevitably leads to the accumulation of formaldehyde within the cell nucleus, in direct vicinity of DNA. This may lead to the extensive DNA damage. Most recently, the lethal consequences of the endogenous formaldehyde formation in the cell nucleus have been experimentally demonstrated for cell deficient in both the formaldehyde digesting enzyme ADHS and the enzyme that repairs DNA damages inflicted by formaldehyde. These results put into the forefront the issue of DNA reactivity with respect to formaldehyde. Our focus here is not the well-known effect of DNA crosslinking by formaldehyde, which is a very important but very rare event, but the formaldehyde’s chemical reaction with amino groups of DNA bases. Previous attempts to fully understand the reactivity of duplex DNA with respect to formaldehyde on the basis of extensive experimental and theoretical studies left a pivotal question unanswered. Although it was convincingly shown that fluctuational openings of base pairs play an important role in the process, there was also a strong indication that a major reaction, hydroxymethylation of the cytosine amino group, proceeds without full base pair openings. Does it mean that formaldehyde attacks the cytosine amino group directly at the bottom of the B-DNA major groove? This issue could not be resolved before. Here we address this question computationally.

**METHODS**

Our mechanistic study is based on solvent mapping, which has originally been developed for the identification of binding hot spots of proteins. The method has been extended to nucleic acids specifically for the present work, and provides a powerful tool for the analysis of interactions between DNA and small organic molecules.

**RESULTS & CONCLUSIONS**

The cytosine amino group, the primary target for formaldehyde reaction in free bases, is totally inaccessible within the double helix. Thus, the duplex DNA chemical modification by formaldehyde can only due to fluctuational violations of the B-DNA structure. It has recently been discovered that Watson-Crick (WC) base pairs in B-DNA spontaneously flip into Hoogsteen (HG) pairs with a reasonably high probability. We therefore performed the computational analysis of the effect of Hoogsteen pairs on the accessibility of cytosine amino groups for the formaldehyde reaction. We found that HG pairing increases the accessibility of amino groups of cytosines, participating in WC pairing adjacent to HG pairs, to small molecules, changes the local binding properties of the DNA, and dramatically increases the exposure of the amino groups to formaldehyde reaction. The incorporation of an HG pair into B-DNA results in very substantial clustering of formaldehyde molecules in close vicinity of the amino group nitrogen of a cytosine participating in WC pairing but adjacent to the HG pairs (Figure 1). A substantial portion of the formaldehyde molecules in the cluster has orientation favorable for chemical attack leading to hydroxymethylation of the amino group. These results indicate that some aspects of DNA reaction with formaldehyde can be explained on the basis of HG breathings of base pairs.

**REFERENCES**

Computational Analysis of Glucosaminoglycan Binding to Chemokine CCL5

Arumina Singh1,2*, Deirdre Coombe1 and Robert J. Woods1,2*

1Institute of Bioinformatics1 and Complex Carbohydrate Research Center2, University of Georgia, Athens, GA, School of Biomedical Sciences, Curtin University, Perth, Australia3, School of Chemistry4, National University of Ireland, Galway, University Road, Galway, Ireland, rwoods@ccrc.uga.edu

RANTES or CCL5 is a chemotactic protein that causes migration of leucocytes upon immobilization on glucosaminoglycans, located on the surface of endothelial cells and in the extracellular matrix. This study investigates the molecular basis of RANTES-heparin interactions and the influence on sulfation pattern of heparin upon binding.

Introduction

Chemokines are small proteins (8-10 kDa) that induce migration of normal tissue cells in repair and development during homeostasis (homeostatic chemokines) and migration of immune cells to the site of infection during an immune response (inflammatory chemokines). Pro-inflammatory chemokines get immobilized by binding to the glucosaminoglycan (GAG) moiety of proteoglycans present on the surface of endothelial cells and the extracellular matrix. The immobilized chemokines in turn interact with specific members of the seven trans-membrane G-protein coupled receptors family present on the surface of immune cells for their localization.

In humans approximately 50 chemokines and 20 receptors are known. The tertiary structure consists of a disordered N-terminus followed by a long loop called the N-loop, followed by a 3-10 helix, 3 stranded β-sheet and C-terminal helix [1]. Chemokines are classified, based on the spacing between the first two cysteine residues, into four categories: C, CC, CXC and CXC (where C is the cysteine residue and X/X3 are one or 3 non-cysteine residues). These cysteine residues form characteristic intramolecular disulfide bonds that stabilize the tertiary structure.

RANTES (regulated on activation normal T-cell expressed and secreted) or CCL5 is a CC type chemokine that shows affinity to glucosaminoglycans (GAGs) such as heparin, chondroitin sulfate and dermatan sulfate on the surface of endothelial cells. It has however been shown to have a particularly high affinity for heparin. Upon secretion from endothelial cells and activated leucocytes it gets localized and causes migration of T-cells, monocytes, basophils, eosinophils, natural killer cells and dendritic cells to the site of inflammation.

RANTES-heparin interaction has been examined by mutagenesis studies and shown to be mediated through the highly basic BBXB motif, RKNR^47 motif on the 40s loop on the surface of the protein [2] (Figure 1), but the details of molecular interaction are not fully understood.

Heparin is highly heterogeneous, with repeating disaccharide units of β-D-glucuronic or α-L-iduronic acid and D-glucosamine linked by 1–4 linkages, and sulfated heterogeneously. However, the influence of the position and amount of derivatization on binding affinity towards RANTES is not entirely understood.

In this study, we present computational analysis of interaction between RANTES and various heparin oligo-saccharides using automated docking and molecular dynamics simulations.

Methods

Automated docking was performed using the molecular docking and virtual screening program, Autodock VINA [3] to dock heparin oligosaccharides on RANTES dimer (PDB ID 1U4L [4]). The entire surface of the protein was used as the search space for the docking (blind docking). Selective docked models were subjected to all-atom MD simulations in explicit water, using the PMEMD program from AMBER 11 molecular dynamics simulation package [5]. Protein force-field ff99SB [6] and carbohydrate force-field GLYCAM06 [7], augmented for sulfate groups, were used for the simulations. The solvated systems were energy minimized for 25000 steps, heated to a temperature of 300 K and then simulated at 300 K for 100 ns. A non-bonded cutoff of 8 Å was used throughout the simulation.

Results & Conclusions

Key structural changes in protein conformation have been observed upon heparin binding. Further analysis will help understand the role played by these conformational changes in heparin binding. Binding energy calculations are also being performed to evaluate the contribution of the difference in sulfation pattern of heparin on binding.

References

RESOLVING THE PARADOX OF CASPASE-6 SELF-ACTIVATION: CRYSTALLOGRAPHIC & BIOPHYSICAL INSIGHTS INTO CASPASE-6 REGULATION

Micah Steffek1, Michael Elliott2, Karen Stanger2, Rami Hannoush1, Structural Biology Expression Group1, Tony Giannetti2, Barthelemy Demeule3, Christine Poznian2, Joe Lewcock2 & Jeremy Murray2

Depts. of Structural Biology1, Protein Chemistry2, Early Discovery Biochemistry1, Biochemical Pharmacology4, Process Research and Development5 and Neurodegeneration6, Genentech Inc. *murray.jeremy@gene.com

We present the crystal structure of pro-Caspase-6 zymogen along with biophysical characterization of Caspase-6 using X-ray crystallography, SAXS, MALS, AUC, and SPR. Our studies provide a plausible mechanism for the regulation of Caspase-6 activation. This hypothesis is supported by the identification and characterization of peptides that bind Caspase-6 zymogen and result in tetramerization and inhibit the activity of the protein.

INTRODUCTION

Caspase-6 is an executioner caspase that has been implicated as an important mediator of cell death pathways and has been linked to neurodegenerative diseases including Alzheimer’s and Huntington’s disease. Recent structural and biochemical studies revealed that Caspase-6 has a unique mechanism of self-activation1. Paradoxically, given this self-activation mechanism, large amounts of pro-Caspase-6 zymogen have been identified in different cell types1,2. In this study, we present the first crystal structure of proCaspase-6 zymogen, a new crystal form of ΔproCaspase-6 zymogen and biophysical characterization of Caspase-6 in solution using SAXS, MALS, AUC and SPR.

METHODS

Pro-Caspase-6 C163A zymogen was crystallized at 4°C in hanging drops, suspended over 25% (w/v) PEG200, 0.05 M Lithium sulfate monohydrate, sodium citrate tribasic dihydrate pH 5.0. Δpro-Caspase-6 zymogen (24-293 C163A) was crystallized at 4°C in hanging drops suspended over 45% (w/v) 2-Methyl-2,4-pentanediol, 0.3 M Ammonium Phosphate monobasic, and 0.1 M Tris pH 8.5.

Size and oligomerization state were determined by SEC-MALS. The instrument measures the protein concentration by UV absorbance, the refractive index, and the scattering intensity of the eluted peaks. Samples were chromatographed at 0.5 ml/min in buffer containing either 50 mM Tris pH 8.0, pH 7.4 or NaAcetate pH 5.5, 200 mM NaCl, 5% Glycerol and 5 mM DTT over an S200 Superdex 10/300 GL column. The Rayleigh ratios of the proteins were determined according to established procedures. Analytical ultracentrifugation sedimentation velocity experiments were conducted using a Optima XL-I analytical ultracentrifuge equipped with absorbance optics. All data acquired from these experiments were obtained using the UV/Visible absorbance detection system of the ultracentrifuge and double sector 12-mm charcoal-filled Epon centrepieces. All measurements SPR were done using a BIACORE 3000 in HBS-EP buffer. The purified avi-tagged proteins were immobilized on a streptavidin chip. Untagged Δ pro-domain Caspase-6 (C163A) and active Caspase-6 were used as analyte for the binding analysis. All other Caspases-6 proteins were covalently immobilized through lysine coupling on a CM5 chip. Small angle X-ray scattering experiments were performed at Stanford Synchrotron Radiation Laboratory, Beamline 4-2, using a wavelength = 1.24Å, sample to detector = 1.7 meter and the MarCCD165 detector. Proteins were dialyzed overnight into the same buffers used for SEC-MALS characterization.

RESULTS & CONCLUSIONS

The work presented here explains the paradox of Caspase-6 self-activation by suggesting that homotetramer formation is the mechanism that regulates Caspase-6 self-activation. Our model of Caspase-6 regulation is consistent with studies that demonstrate the presence of reservoirs of proCaspase-6 zymogen in cells and provides a plausible explanation for the regulation of Caspase-6 activity. This work highlights the uniqueness of Caspase-6 compared to other caspases and provides insight into the novel mechanisms regulating Caspase-6 that may be useful to understanding the biological role of Caspase-6 and for the development of therapeutics for diseases that involve Caspase-6. (Figure 1).
**INTRODUCTION**

p53 is a transcription factor involved in expression of a number of downstream genes in response to genotoxic stress. In normal cells, it is present in the latent or inactive state and in case of cancer cells it is activated by various post translation modifications. It is found to be mutated in 50% of the cancers. These mutations occur at a high frequency in the DNA binding region of the p53. The DNA binding region comprises of loop-sheet-helix motif involved in major groove binding and two large loops and Zinc (Zn) ion involved in minor groove binding. Amongst the seven known hot spot cancer mutations G245S, R249S and R273C have been studied here using molecular dynamics (MD) simulations. These mutations are experimentally proven rescue mutations which have also been included in the present work. A comparative study of these cancer mutations along with wild type and their rescue mutations has been presented. A 30 ns simulation study was analysed to observe the local structural changes and DNA binding property of p53 in case of wild type, cancer and rescue mutants.

**METHODS**

The simulations were performed using the AMBER 10 suite of programs. The double stranded DNA, Zn ion and chain B of 1TSR PDB structure was considered as the starting conformation. In order to obtain the structure for mutants 1TSR (chain B) was mutated and rebuilt in the *xleap* module of AMBER. A total of seven simulations were performed which included wild type, three cancer mutants (G245S, R249S and R273C) and their corresponding rescue mutants (G245S_H178Y, R249S_T123A_H168R and R273C_T284Y). The p53-DNA complex was further neutralized by Na+ ions and solvated using the TIP3P water model. The Zinc ion is present in tetrahedral co-ordination with four residues of p53. This co-ordination complex was maintained using the quantum mechanical method, PM3. The parameters for Zn were obtained from the work performed by Lu Quang et. al. The remaining system was simulated using the all atom point charge force field FF03 from AMBER 10. Initial minimization of 20000 steps was performed. Temperature was increased to 300 K using the Langevin dynamics and 2 fs time step. An equilibration simulation for 2 ns was performed. A production run for 30 ns for each of the seven simulations was performed leading to a total of 210 ns.

**RESULTS & CONCLUSIONS**

Each trajectory of 30 ns simulation was analyzed at every 10 ps snapshot. Each of the mutant leads to some localized conformational change which can be observed by calculating the Root Mean Square Deviation (RMSD) against the crystal structure (1TSR). In case of G245S and R249S, experiments suggest conformational changes in the loop regions of the major and minor groove binding sites. This was confirmed by an increased value of RMSD in case of the cancer mutants as compared to that of wild type and rescue mutants. R273C, structurally behaves similar to that of wild type but leads to loss of DNA binding. The DNA binding property of each of the cancer and rescue mutants were defined by comparing the hydrogen bonding between p53 and DNA with free energy of binding ($\Delta G$). These two properties are negatively correlated as a decrease in the hydrogen bond formation leads to loss of binding which increases the free energy of binding. The quantitative information of this correlation observed in each of the cancer and rescue mutants has been calculated. The cancer mutants lead to drop in the number of hydrogen bonds (NOH) causing an increase in the free energy of binding. In case of rescue mutants there was gain in NOH leading to decrease in free energy, This work helps to conclude that by calculating and correlating the free energy of binding and NOH in various p53 mutants would help to differentiate between its cancer or rescue behavior.

**REFERENCES**

We developed the SymD program to find internally symmetric protein domains and noticed some with more than one internal symmetry element. Here we systematically identified 29 domains among 10,569 domains in ASTRAL 1.75. 40% sequence identity set in order to learn the types of possible symmetry combinations and the architecture of the protein structures that allow such symmetry combinations.

**RESULTS & CONCLUSIONS**

The number of proteins with multiple symmetry axes at domain level is very different from that at the oligomeric complex level. The number of symmetric proteins found, using the new version of SymD, is 21% of all known domains, up from 10 to 15% estimated earlier using the original version of SymD. However, the number of proteins with multiple non-slip symmetry elements is only 29, which is 1% of symmetric proteins and a mere 0.3% of all domains. This is in stark contrast to the number of non-cyclic symmetric protein complexes, which make up 26% of all symmetric homomeric complexes. Similar to the functional relation found in symmetric protein complex, many domains with the internal dihedral symmetry are part of an enzyme or chaperon and many with bent helical and superhelical structures are involved in cytoskeleton and interactions with membrane in which directionality is important.

**REFERENCES**


**CONCLUSIONS**

The number of proteins with multiple symmetry axes at domain level is very different from that at the oligomeric complex level. The number of symmetric proteins found, using the new version of SymD, is 21% of all known domains, up from 10 to 15% estimated earlier using the original version of SymD. However, the number of proteins with multiple non-slip symmetry elements is only 29, which is 1% of symmetric proteins and a mere 0.3% of all domains. This is in stark contrast to the number of non-cyclic symmetric protein complexes, which make up 26% of all symmetric homomeric complexes. Similar to the functional relation found in symmetric protein complex, many domains with the internal dihedral symmetry are part of an enzyme or chaperon and many with bent helical and superhelical structures are involved in cytoskeleton and interactions with membrane in which directionality is important.

**REFERENCES**

THE ANALYSIS OF INTERACTIONS BETWEEN GALK ENZYME AND KNOWN INHIBITORS SUGGESTS GUIDELINES TO DEVELOP NEW DRUGS FOR GALACTOSEMIC PATIENTS

Federica Chiappori, Ivan Merelli, Luciano Milanesi & Anna Marabotti*  
ITB-CNR, Italy. *anna.marabotti@itb.cnr.it

Interactions between GALK enzyme and known inhibitors have been studied by computational biology approaches such as docking and molecular dynamics simulations. Results allowed us to identify the most persistent anchoring points between the inhibitors and the enzyme, to find a novel putative binding cavity and to collect information for the development of a new generation of GALK inhibitors with an increased specificity and affinity.

INTRODUCTION

Classic Galactosemia is a rare genetic disease caused by the impairment of the enzyme galactose-1-phosphate uridylyltransferase (GALT) involved in the second step in the galactose metabolism (Scheme 1).

Accumulation of galactose-1-phosphate (gal-1-P) is regarded as the main cause of the chronic symptoms of galactosemic patients. Blocking the activity of galactokinase (GALK) would hamper the formation of this molecule, causing only mild symptoms, similar to those of people with another form of galactosemia, namely galactosemia type 2 (Scheme 1). Several efforts have been made in the past to develop molecules able to bind GALK and inhibit its activity, but at present, only few inhibitors with moderate affinity and selectivity have been identified with several high-throughput screenings1.

In an effort to identify new inhibitors of GALK enzyme with an improved affinity and selectivity, we used computational approaches such as docking and molecular dynamics (MD) simulations in order to study the molecular details of the interaction between the enzyme and its inhibitors. The information derived from our results can be used in rational drug design studies to identify ligands with improved potency, or selectivity, or both, resulting as promising candidates for drug development for galactosemic patients.

METHODS

The X-ray structure of GALK enzyme complexed to galactose and an ATP analogue (PDB code: 1WUU)2 was used as starting point for the simulations. Three-dimensional structures of known ligands were taken from PubChem database3. Docking simulations were performed using AutoDock 4.2.3 and preparing the structures with ADT 1.5.2.4 MD simulations were performed with Gromacs 4.5.35, using the force field Amber99SB6, setting a rhombic dodecahedric box filled with water and NaCl7 ions. After stabilization with short simulations using NVT and NPT ensembles, the final MD simulations were run for 30 ns at 310K and 1 bar constant pressure. Analyses were conducted using the software from Gromacs package and other in-house Perl scripts.

RESULTS & CONCLUSIONS

Our computational approach allowed to dissect at molecular level the interactions between GALK enzyme, its natural substrate galactose and the set of inhibitors found in previous high-throughput screenings. In addition to the canonical GALK binding site, we identified a secondary binding cavity in which galactose and most of the ligands are predicted to bind. This cavity could be of interest in the perspective of designing ligands with increased specificity and/or affinity for the enzyme. The interactions of GALK with known inhibitors are rather different compared to natural substrate. Indeed, galactose interacts mainly by H-bonds in the binding sites, with a lower contribution of hydrophobic contacts, whereas the opposite is true in the case of the analyzed ligands. This could be the cause of their low affinity towards the enzyme. In addition, we identified those residues of the protein capable of establishing the most stable interactions with galactose and selected inhibitors, and we analyzed the evolution of these interactions during the time.

This knowledge would improve the development of a new class of GALK inhibitors, in which e.g. the introduction of hydrophilic moieties to increase the strength of interaction with the enzyme could be directed only to those moieties that actually can interact with high efficacy. Also the knowledge of those residues in the protein involved in other kind of contacts (such as hydrophobic contacts and stacking interactions) could enhance the affinity of the inhibitors. Finally, the identification of those residues responsible for the selectivity towards other enzymes of the kinase family would improve the development of more specific ligands.

REFERENCES


ACKNOWLEDGMENTS

PGC Research Grant, FIRB ITALBIOMET project (RBPR05ZKZZ and RBIN064YAT_003), Flagship Project InterOmics.
DEVELOPMENT OF A MULTISCALE FRAMEWORK TO STUDY THE ENERGETICS OF PROTEIN DYNAMICS IN PROTEIN-PROTEIN INTERACTIONS

Nils J. D. Drechsel1, César L. Ávila2, Michael A. Johnston3, Jordi Villà-Freixa4*

Universitat Pompeu Fabra1 Universidad Nacional de Tucumán2, University College Dublin3.

*jordi.villa@upf.edu

A multiscale framework is presented that enables a fast sampling of the conformational space of proteins using a physics-based coarse-grained force-field which is based and made compatible to the Amber force-field series. The coarse-grained energies are in good agreement with all-atom Amber energies leading to a similar exploration of the respective conformational spaces. Deviating energetics are corrected with a multiscale free energy perturbation.

INTRODUCTION

State of the art supercomputers, specifically designed to simulate the dynamics of molecules, are nowadays capable of calculating the motions of small solvated proteins over a temporal range of milliseconds. They provide detail both in high temporal and spatial resolution. The costs of these productions however is unnecessarily massive, since such methodology represents a brute-force attempt, and, depending on the targeted question, might be produced with lean and significantly faster methods. One such method is the focus of this poster. In the presented multiscale framework, we show how the sampling problem of free energy calculations is tackled. This approach uses a thermodynamic cycle in which, instead of evaluating directly the free energies of a reaction, the same is performed in low resolution using a coarse-grained force-field. Unavoidable disagreements due to the coarse-grained nature of the force-field are corrected with a free energy perturbation, in which the difference in free energy of going from low detail to high detail is determined. The coarse-grained force-field that is used here is called AmberCG. It is functionally based on Warshel’s coarse-grained force-field, but was further modified to be compatible to the Amber force-field series by a reparametrization obtained from a genetic algorithm. The resulting parametrization was validated against a databank of 21 proteins, representing all prominent folds.

METHODS

Free energy perturbation is a standard technique to calculate the difference in free energy between compatible systems of interest, usually used to compute alchemical pathways which lead to vast performance improvements compared to their pathway complements. To study the effect of mutations on protein-protein interactions, the change in complex stability, as well as folding stability needs to be determined - expensive in the physical pathway, cheap in the alchemical pathway, even cheaper in a multiscale pathway. In the alchemical pathway, we calculate differences in free energy between the wt and mutant unfolded states, the wt and mutant folded but unbound states and the wt and mutant bound states respectively (see figure 1). Using an additional multiscale pathway, these computations are performed using the AmberCG coarse-grained force-field, where the additional correction for the costs of moving from the all-atom to the coarse grained potentials needs not be calculated because it is canceled out in the double difference.

The AmberCG force-field is a coarse-grained force-field in which the side-chains are replaced by cg-balls, the main-chain atoms most notably are not modified. It is compatible to the Amber series, since it uses similar potentials for its main-chain interactions. The interactions were fine-tuned using a genetic algorithm in which the correlation between coarse-grained energetics and all-atom energetics were improved for a database of 20000 decoy structures.

RESULTS & CONCLUSIONS

The optimization was performed on the La Palma supercomputer for a total of approximately 100K individual solutions, resulting in correlations to the decoy structures of 0.91. The parameterization was validated using a set of 21 trajectories of different structures representing all major folds, which were simulated for 10ns each. Some of the structures correlated well, the highest with 0.97, while others correlated moderately, the lowest with 0.53, giving an average correlation of 0.74.

REFERENCES

STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF MUTANTS OF GALT ENZYME

Anna Marabotti2, Angelo Facchiano2, Luciano Milanesi2, Manshu Tang1 & Kent Lai1
ITB-CNR, Italy1, ISA-CNR, Italy2 and Dept. of Pediatrics, University of Utah School of Medicine, USA2.
a@itb.cnr.it and kent.lai@hsc.utah.edu

Computational biology approaches were applied to characterize at molecular level the effects of mutations found on GALT enzyme in galactosemic patients. Results show that in many cases these mutations alter the structural features and the stability of the protein, and this has effects on its functions. These results are in agreement with experimental characterization of these mutants.

INTRODUCTION
The enzyme galactose-1-phosphate uridylytransferase (GALT) catalyses the conversion of galactose-1-phosphate to UDP-galactose, the second step in the galactose metabolism. Mutations in the GALT gene are associated to a rare genetic disease called Classic Galactosemia that can cause severe symptoms to affected people. To date, only few studies were made about the biochemical characterization of the impact of these mutations on GALT structure and function1. A previous study has allowed to predict the effect of known mutations by computational methods2 and a database has been created to store all these results at: http://bioinformatica.isa.cnr.it/GALT3

In this work, we have characterized the effects of 14 novel missense mutations on structure and function of GALT enzyme, using a computational biology approach. Results were compared to the biochemical characterization of these mutants, expressed and purified from bacteria.

METHODS
Starting from the 3D model of human GALT enzyme4 homodimeric mutants were created using Modeller 9v85. Each resulting mutant was analysed for variations in secondary structures, solvent accessibility, intersubunit interactions, H-bond and salt bridge patterns, predicted stability of the protein, as it was made previously6. In addition, selected mutants for which the static modelling of mutations did not allow to highlight any variation in these properties, were submitted to molecular dynamics simulations using GROMACS program6, and results were compared with those obtained on wild type enzyme and on the most characterized mutant, Q188R.

The human homodimeric mutant proteins were expressed in E. coli, purified and assayed for their activity and kinetic properties (Vmax and Kcat).

RESULTS & CONCLUSIONS
The analysis of the models of homodimeric mutant GALT enzyme showed that in most cases these proteins have stability problems. At structural level, these mutants show in two cases the alteration of intersubunit interactions, in other two cases the involvement of residues indirectly affecting the active site, and in most cases the alteration of the network of H-bonds and salt bridges, with either local or global effects.

The results of molecular dynamics simulations for those cases where static modelling was not able to predict any structural effect, were able to suggest possible effects that impair the correct activity and the stability of the enzyme, too. The comparison with simulations made on mutant Q188R allowed also to confirm the deleterious effects of this mutation at structural level.

The biochemical characterization of these mutants showed results that are well explained by the simulations. In particular, those mutations that were predicted to have a large impact on protein structure are also those with minimal or no activity, whereas those mutants showing a residual activity, also show localized effects on protein structure.

Therefore, the combined use of the “wet” and “in silico” techniques offers novel and important insights into mutational analysis of the human GALT gene.
EFFICIENT MODELING OF PEPTIDE FOLDING, ORIENTATION,
AND ASSEMBLY IN LIPID BILAYERS USING IMPLICIT SOLVATION

Andrew J. Bordner, Barry Zorman, and Ruben Abagyan

1Mayo Clinic, Scottsdale, AZ; 2Molsoft LLC, San Diego, CA. bordner.andrew@mayo.edu

Membrane proteins function within the heterogeneous environment of solvated lipid bilayers. We have developed an implicit solvation model to efficiently account for this environment in all-atom molecular modeling. Despite the simplicity of the model, computational predictions of peptide folding, orientation, and association agree well with diverse experimental measurements.

INTRODUCTION
Integral membrane proteins constitute a significant portion, typically 20-30%, of the proteomes in sequenced organisms. They are also the targets of approximately half of approved drugs. Despite their ubiquity and biomedical importance relatively few experimental structures of membrane proteins are available due to technical challenges. Computational methods can address this problem by providing atomic resolution structures that can be used to guide experiments and aid in drug discovery.

Membrane proteins reside within the spatially and chemically heterogeneous environment created by a water solvated lipid bilayer. Molecular dynamics simulations with explicit solvent and lipid molecules provide the most detailed solvation model but are computationally expensive. In contrast, we have parameterized an implicit solvation model in which the average effects of interactions with lipids and solvent are efficiently calculated using atomic solvation parameters (ASPs)1.

METHODS
The solvation energy is calculated using ASPs that are functions of both the atom type and membrane normal coordinate. The ASPs in the membrane hydrophobic core were fit to gas-trans-cyclohexane transfer free energies for residue side chain analogs while previous values were used for ASPs in the aqueous solvent2. ASPs in the membrane interface were calculated by linear interpolation between these two values. The optimal atom types were determined in an unbiased manner by clustering and ASPs fit by linear regression. Furthermore, the optimal membrane geometry was determined by minimizing the error between the predicted orientation of five transmembrane (TM) peptides and experimental solid state NMR data.

Molecular mechanics simulations with the implicit solvation model were performed using the ICM program (Molsoft LLC). This program uses biased-probability Monte Carlo sampling in internal coordinate space to search for the global minimum of a physical energy function (ECEPP/3)1.

RESULTS & CONCLUSIONS
The optimal membrane geometry was 30 Å thick with a 20 Å central hydrophobic region. The error in the predicted tilt angle of the TM peptides was less that 5° for all peptides except one outlier (PDB entry 1MZT) with an error of 17.7°

Next, we performed ab initio folding of six different amphipathic peptides starting from a fully extended conformation. The folded peptides were properly localized to the solvent interface region and their conformations agreed with NMR structures, including one peptide (cecropin A-magainin 2 hybrid) with two alternative conformations.

In addition, we predicted the structures for a series of 20 host-guest peptides in both aqueous and non-polar solvent environments and compared the structures with the experimental results of Liu and Deber3. We found that the predicted helicities were significantly correlated for the non-polar solvent (membrane core) but not for the aqueous environment. We also predicted the secondary structures of several homopolypeptides in both non-polar and aqueous environments and analyzed the results in the context of experimental and theoretical secondary structure propensities.

Wimley et al. examined a series of host-guest peptides, acetyl-WLLXLL-OH, and found that some aggregate into β-sheets in membranes4. Simulations of dimers using the solvation model yielded good agreement with experimentally measured secondary structure propensities.

REFERENCES
SELECTING THREE-DIMENSIONAL TRANSMEMBRANE BETA BARREL MODELS USING PREDICTED Z-COORDINATES

Sikander Hayat & Arne Elofsson
Center for Biomembrane Research, Department of Biochemistry and Biophysics,
Stockholm Bioinformatics Center, Science for Life Laboratory, Swedish E-science Research Center, Stockholm University, SE-10691 Stockholm, Sweden. email@bioinfo.se

We present a computational method called tobmodel for generating 3D models of transmembrane β-barrels. First, alternative topologies are obtained from the BOCTOPUS topology predictor. Then several 3D models are constructed by using different tilt of β-sheets. Finally, the best model is selected based on agreement with a novel predictor, ZPRED3, which predicts the distance from the center of the membrane for each residue, i.e. the Z-coordinate. The Z-coordinate prediction has an average error of 1.61 Å. Tobmodel predicts the correct topology for 75% of the proteins in the dataset. Importantly, tobmodel provides a Cα template with an average RMSD of 7.24 Å from the native structure. Tobmodel is freely available as a web server at: http://tobmodel.cbr.su.se

INTRODUCTION
Transmembrane β-barrels exist in the outer membrane of gram-negative bacteria as well as in chloroplast and mitochondria. They play a major role in transport processes and are promising antimicrobial drug targets. Since structures of only a few transmembrane β-barrel families are known, a computational method for their 3D modeling would be valuable. From available transmembrane β-barrel 3D structures we know that all bacterial transmembrane β-barrels have an even number of anti-parallel transmembrane β-strands. This symmetrical arrangement of the barrels suggests that an approach based on idealized geometries may be successful.

METHODS
The pipeline for model generation and ranking is shown in Figure 1. As shown, BOCTOPUS [1] is used to obtain alternative topologies for the given sequence. Then, multiple Cα models of the transmembrane β-barrel region are generated for different tilts of β-strands for all obtained topologies [2, 4, 5]. A novel z-coordinate predictor called ZPRED3 is used to predict the distance of residues from the membrane center. The top-ranking model is then chosen based on the minimum difference between the predicted z-coordinate and the z-coordinate obtained from the generated models [2].

RESULTS & CONCLUSIONS

<table>
<thead>
<tr>
<th>Method</th>
<th>Correct Topology</th>
<th>Average RMSD</th>
<th>Average TM_Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMBpro</td>
<td>19(36)</td>
<td>8.79</td>
<td>0.56</td>
</tr>
<tr>
<td>tobmodel</td>
<td>27(36)</td>
<td>7.24</td>
<td>0.43</td>
</tr>
<tr>
<td>BOCTOPUS</td>
<td>25.4±2.0(36)</td>
<td>NA*</td>
<td>NA*</td>
</tr>
<tr>
<td>3D-SPoT</td>
<td>NA(23)</td>
<td>4.10</td>
<td>0.5</td>
</tr>
<tr>
<td>tobmodel</td>
<td>18(23)</td>
<td>7.06</td>
<td>0.43</td>
</tr>
<tr>
<td>tobmodel</td>
<td>23(23)</td>
<td>5.86</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table 1. Comparison of models generated by different prediction methods. TMBpro employs pre-defined templates extracted from proteins in their dataset. *tobmodel dataset. **BOCTOPUS does not generate three-dimensional models. ***3D-SPoT dataset. **3D-SPoT does not test alternative topologies and the results are taken as reported previously [3]. For comparison, here we have used the correct topologies into the tobmodel pipeline.

Table 1 shows the results obtained based on a 10-fold cross-validation. As shown, tobmodel selects the correct topology for 27 out of 36 proteins in the dataset. TMBpro [6] and BOCTOPUS predict the correct topology for 19 and 25.4±2.0 proteins, respectively. Models obtained from TMBpro and tobmodel have an average RMSD of 8.79 and 7.24 Å.

The average TM_Score for TMBpro models is 0.56 and is slightly higher than for top-ranking tobmodel models (0.43). Top-ranking models obtained from tobmodel are also compared with models obtained from 3D-SPoT [3]. The lower three rows in table 1 show that the RMSD for 3D-SPoT models is slightly better than models obtained from tobmodel. However, we believe that tobmodel can be a useful tool for topology prediction and 3D modeling of transmembrane β-barrels. In future, more advanced model selection methods will be developed to select the best possible model.

REFERENCES
APPLICATION OF EXHAUSTIVE PROTEIN-PROTEIN INTERACTION PREDICTION SYSTEM BY USING PROTEIN DOCKING TO SIGNAL TRANSDUCTION PATHWAYS

Yuri Matsuzaki¹, Nobuyuki Uchikoga², Masahito Ohue² & Yutaka Akiyama¹
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology¹, Department of Physics, Chuo University². y_matsuzaki@bi.cs.titech.ac.jp

We have developed a high throughput protein-protein interaction (PPI) prediction system “MEGADOCK”, which is based on rigid-body docking. It is designed to be suitable for running on massively parallel computing environments. MEGADOCK showed significant speed-up in rigid-body docking process that led to a way of full utilization of the proteins tertiary structure data for large-scale, network-level problems. In addition to evaluate our PPI prediction system using a benchmark dataset, we applied our system to reconstruct known signal transduction systems.

INTRODUCTION

We are working on developing a computational method of protein-protein interaction (PPI) prediction, one of the important problems of systems biology. Usually, the expected role of the computational method using physical chemistry on PPI analysis was to examine the configuration and affinity of bonds concerning the known one-to-one protein-protein interaction in detail. Here we explored further utilization of rigid docking method by applying it to predictions of possible binding protein pairs from a large amount of protein groups. This system is called “MEGADOCK,” which predicts candidate pairs of PPI by exhaustive docking and post-docking analysis.

METHODS

MEGADOCK is a system to predict the presence or absence of interaction using information about the tertiary structure of proteins based on various scores obtained from rigid-body docking. In this calculation, a high-speed evaluation is conducted mainly based on the shape complementarity of the molecular surface without considering the structural change of the protein. We introduced the rPSC (real Pairwise Shape Complementarity) score composed of the terms of shape complementarity and electrostatic interaction assigned to the molecular structures on the voxel space.

MEGADOCK is parallelized using OpenMP and MPI library. Upon docking, ligand protein structure is rotated to 3600 different angles and then the system conducts a translational search of favourable binding site. Calculations of 3600 angles are distributed by thread parallelization implemented with OpenMP. Because the calculations for each pair are independent, we can parallelize an all-to-all exhaustive PPI prediction task on hundreds or thousands of CPU cores. The user can specify the number of receptor and ligand protein data to be assigned to a single processor after taking the memory capacity into consideration. We have tested this data parallelization using up to 12,288 nodes and showed that it is scalable on large number of nodes (Figure 1).

Prediction of the relevant PPIs is performed according to the affinity scores calculated by the post-processing of all the docking results. Each docking calculation outputs several thousands of decoys that yield high docking score. We defined affinity scores based on the highest docking score and distribution of the docking scores of the high scoring decoys. It represents how improbable to find that highest score among the score distribution if the given pair of proteins is just a random pair that do not actually interact.

RESULTS & CONCLUSIONS

As an application, PPI prediction using MEGADOCK was performed on the reconstruction of a canonical signal transduction pathway of bacterial chemotaxis (13 proteins, collected 89 structure data including multiple structures for each protein species) and human EGFR signal transduction pathway (49 proteins, 497 structures). The F-measure value was 0.464 when applied to chemotaxis system and 0.385 when applied to human EGFR system.

The proposed approach to computational PPI detection is a promising methodology for mediating between structural studies and systems biology by utilizing cumulative protein structure data for pathway analysis.

REFERENCES


ACKNOWLEDGEMENTS

Part of the result is obtained by early access to the K computer at the RIKEN Advanced Institute for Computational Science.
PREDICTION AND ANALYSIS OF IDIOTYPE-ANTI-IDIOTYPE ANTIBODY COMPLEXES ASSOCIATED TO CELIAC DISEASE

Anna Vangone1, Daniele Sblattero2, Ivana Caputo1, Luigi Cavalli1 & Romina Oliva1,2
1. Dept. of Chemistry and Biology, University of Salerno, Italy 2. Dept. of Health Sciences, University of Piem. Orient., Italy 1. Dept. of Applied Sciences, University of Naples “Parthenope”, Italy. oliva@uniparthenope.it

We performed docking simulations between the celiac clone2.8 antibody and two anti-idiotypic antibodies it elicited in human and mouse. Multiple docking solutions were analyzed based on intermolecular contact maps, and compared with experimental structures of anti-idiotypic antibody complexes. Results clearly indicate that human and mouse anti-idiotypic antibodies dock to clone2.8 with a different binding mode.

INTRODUCTION
Antibodies, the main actors of immune response, can themselves be antigenic under a variety of circumstances. An antibody (Ab2) elicited by another antibody (Ab1) and directed against its antigenic-determinants is known as an anti-idiotypic antibody. Recently, celiac antibodies recognizing tissue transglutaminase, the main celiac auto-antigen, have also been shown to elicit the production of anti-idiotypic antibodies in human and mouse.1 Importantly, anti-idiotypic antibodies may bring an ‘internal image’ of the original antigen, and this is at the basis of the production of new generation vaccines.2 However, in spite of its valuable interest, the mutual binding of antibodies is to date under-investigated from a structural point of view, as only few experimental structures of Ab1-Ab2 complexes are currently available from the PDB.

Herein we present the results of docking simulations between clone2.8, a prototype celiac antibody, and two anti-idiotypic antibodies, AImV1 from human, and Aft2 from mouse. We also compare the obtained solutions with the experimental Ab1-Ab2 structures available from the PDB.

METHODS
The variable domain structures of clone2.8, Ab2-human and Ab2-mouse were modeled by the RosettaAntibody Fv homology modeling server,3 using the full refinement protocol option. The obtained models were then used for protein-protein docking simulations, performed by the ClusPro 2.0 server.4 In all the simulations, all the Ab2-human and Ab2-mouse residues that do not fall into the Complementary Determining Region (CDR) were masked (Cluspro Antibody Mode). Differently, for clone2.8 two situations were explored. In the former, indicated as ‘blind’ docking, all the clone 2.8 residues were considered on an equal basis; in the latter, indicated as ‘active’ docking, the CDR residues were marked as attractive, to have them interacting. All the analyses and visualizations of the docking solutions and crystallographic complexes were performed with the COCOMAPS5 and CONS-COCOMAPS6 web tools.

RESULTS & CONCLUSIONS
The ‘blind’ docking simulations gave solutions with relatively low population for both the clone2.8/Ab2-human and clone2.8/Ab2-mouse systems (72 and 86, respectively, for the first cluster). Analysis of the contact map5 of the representatives of the ten best solutions for the two systems indicated that residues of the clone2.8 CDR, although not assigned as attractive, were always involved in the interaction. We then applied CONS-COCOMAPS to the ten best solutions to obtain: i) a ‘consensus map’, i.e. a map where inter-molecular contacts are shown in a scale of gray where the more conserved the contact, the darker the spot, and ii) a list of the most conserved contacts. Overall the ‘blind’ simulations clearly show that the region recognized by the anti-idiotypic antibodies, especially from mouse, involves the clone2.8 CDR loops. We therefore ran ‘active’ docking simulations, where the CDR residues of clone2.8 were selected as attractive. From the consensus maps it is apparent that the ‘active’ solutions represent a subset of the ‘blind’ solutions. Population of the first clusters is about doubled and one preferred solution per each system clearly emerges, that also coincides with the first most populated ‘blind’ solution for the clone2.8/Ab2-mouse system and with the second ‘blind’ solution for the clone2.8/Ab2-human system. In the clone2.8/Ab2-mouse predicted complex all the six hyper-variable loops are involved into the interaction (Figure 1), maximizing the interface area. As for Ab2-human, it possibly binds to the clone2.8 CDR mainly with its L2, H1 and H3 hyper-variable loops, although a different binding mode, not centered on the clone2.8 CDR and corresponding to the first ‘blind’ solution, cannot be excluded.

Therefore, our simulations indicate that the two anti-idiotypic antibodies elicited by clone2.8 bind to it in a different way. Interestingly the preferred docking solution found for the clone2.8/Ab2-mouse system closely resembles the most recurrent binding mode observed in available experimental structures of Ab1-Ab2 complexes.

REFERENCES

FIGURE 1. Representations of clone2.8/Ab2mouse 3D model: Left: COCOMAPS “distance range contact map”. Labels have been added for the antibody hypervariable loops L1-L3 and H1-H3. Red, yellow, green, and blue indicate contacts within 7 Å, 10 Å, 13 Å, and 16 Å, respectively. Right: A Pymol (www.pymol.org) visualization of the complex based on the automatic COCOMAPS script.pml; residues at the interface are shown as ‘sticks’.
A COMPREHENSIVE COMPARISON OF PROTEIN STRUCTURAL ALIGNMENT ALGORITHMS

Lazaros Mavridis\(^1\), Vishwesh Venkatraman\(^2\), & David W. Ritchie\(^2\)

\(^1\)Dept. of Chemistry, St Andrews University; \(^2\)INRIA Nancy, \(^{3}\)m213@st-andrews.ac.uk

The question of how to compare and classify the 3D structures of proteins is one of the most important unsolved problems in computational biology. Here, we compare 17 protein structure-based alignment methods, which cover the current state of the art, based on their performance to retrieve homologues protein domains according to the CATH classification database.

INTRODUCTION

In recent years many structural alignment algorithms have been developed and described\(^3\). The authors of each new algorithm often compare their method to a selection of existing methods using tasks such as pairwise structural alignment on hard protein pairs, and protein domain classification using manually curated classification databases. In the past few years there have been some general reviews\(^4\) on structural alignment algorithms, but there has not been a detailed comparison since Kolodny and Levitt\(^5\).

Here, we present a thorough comparison of 17 popular protein structure-based alignment methods which work reliably on the whole of the CATH\(^6\) database. The ability of those methods to retrieve homologous protein domains is evaluated using a large number of different queries on CATH. We compare the results using several statistical techniques, such as receiver-operator characteristic (ROC) plots and analysis of variance (ANOVA) tests.

METHODS

The following alignment methods were compared: 3DDBLAST, 3DZERNIKE, CE, COMSUBSTRUCT, CPSARST, DALI, FRTMALIGN, LOVOALIGN, MAMMOTH, MATALIGN, MATT, SABERTOOTH, STRUCTAL, SSM, TMALIGN, TRIANGLEMATCH and YAKUSA. These methods were evaluated using the CATH database comprising over 10,000 protein domain structures distributed over 2,549 families. The purpose of the experiment was to test how well each method can retrieve similar domains to a given structure. A large number of queries were selected from the database based on a few basic criteria, which were used so that we can ensure that no biased and artificial enrichments are introduced in the results.

RESULTS

For each alignment method each query was run against the CATH database and we retrieved a ranked list. Given a query the members of the superfamilies that the query is a member of, were marked as positives and all the rest as negatives, using that classification we calculated the corresponding ROC plots. In order to summarize the results and the performance of each of the methods we calculated for each of the methods the aggregate ROC plot by combining the 217 ROC plots per method, as shown in Figure 1. In order to highlight the early recognition performance of the methods tested we also report the aggregate AUC values for the top 5% and 10% of the ranking lists.

We can see from Figure 1 that most methods perform very well with a small number being almost able to reproduce the CATH classification. On the other hand, three methods gave very poor results.

Finally an ANOVA test based on the 217 individual queries was used to compare the performance of the methods based on statistical significance. The result of the test shows that the average AUC values differ significantly (p-value=2e\(^{-9}\)<0.05). The methods SSM, MATALIGN, FRTMALIGN and TMALIGN have the best overall performance with an aggregate AUC of 0.978 (p-value=0.05), followed by 3DDBLAST, MAMMOTH and SABERTOOTH with an aggregate AUC of 0.956 (p-value=0.42). Finally STRUCTAL, LOVOALIGN, YAKUSA and DALI gave aggregate AUC values of 0.9 (p-value=0.57). Although the best methods usually perform very well for the majority of the queries, some queries are difficult for all methods. This probably suggests that some CATH families should be reclassified.

CONCLUSIONS

We have seen big improvements since the early days of DALI and CE (1993-2000). Some of the methods developed during 2000-2005 such as SSM, TMALIGN, and MATALIGN perform significantly better than DALI and CE. Several more recent methods such as FRTMALIGN and 3DDBLAST also perform very well, but despite using more sophisticated algorithms, any improvement has been only rather modest.

REFERENCES

A SEMI-AUTOMATIC PROCEDURE FOR THE COMPARISON OF CONFORMATIONAL AND FUNCTIONAL PROPERTIES OF PROTEIN MUTANTS

Angelo Facchiano1*, Anna Marabotti2*

Institute of Food Science1 and Institute of Biomedical Technologies2, National Research Council, Italy.

* angelo.facchiano@isa.cnr.it

Point mutations in proteins are observed as natural variants in genetic pathologies, or may be designed in protein engineering studies. Their effects at structural and functional levels are of interest in order to understand molecular bases of pathologies or to obtain suitable new features. We developed a semi-automatic procedure for generating molecular models of protein mutants and to evaluate the structural consequences of the mutation, in terms of alteration of secondary structure, exposure and interface interaction, as well as loss or gain of stability, H-bonds and salt bridges, and for functional aspects. The results are presented as html pages, which can be easily used for publishing on the web.

INTRODUCTION

Studies on the effects of natural genetic variations, as well as protein engineering studies, imply the comparison of structural and functional properties of the “wild-type” protein to the mutated form. The comparison can be performed at different levels of investigation, including experimental studies and computational analyses of the 3D models. Each protein presents peculiar features in terms of function, therefore this aspect appears difficult to be analyzed by automatic tools, whereas structural properties can be considered more general features, and automatic procedures can be adopted to observe the effects of mutations on the protein structure, also for large number of mutations. We developed a semi-automatic procedure to investigate both structural and functional effects of point mutations, with the aim of investigate specific proteins of interest in human pathologies.

METHODS

The availability of a 3D model of the protein of interest is the main information needed. An experimental model as well as a predicted model can be used for the study, taking into account that the reliability of the analysis depends on the quality of the starting model. Any additional information in terms of structure and function of the protein, represents an useful knowledge to be integrated into the analysis. Mutations are simulated by using mutate_model script of Modeller. The models of wild-type protein and each mutant are analyzed by many programs: DSSP for the secondary structure analysis, NACCESS for solvent exposure, integrated by an original analysis to determine interface residues in case of oligomeric proteins, HBPLUS for H-bonds, a combination of online tools for salt bridges and detection of residues within a threshold distance from specific atoms of interest for function. Original scripts in Perl have been written to make an automatic extraction of information of interest from the output generated by the programs, and to generate results as html pages. These are created on the basis of a template in which structural information are added. This means that different templates can be used, without differences in the whole procedures, to obtain a customized visualization of results.

RESULTS & CONCLUSIONS

We have investigated in the recent years the role of mutations on GALT, an enzyme involved in classical galactosemia14. More than one hundred point mutations have been modeled and investigated for their effects on the enzyme structure and function. The procedure has been improved to include some more analysis and it has been framed into a semi-automatic tool which can be adapted for using it on other proteins of interest. An example of its application is visible into the GALT db online, at the URL http://bioinformatica.isa.cnr.it/GALT/. The starting model for this study was the model of human GALT3. After the main pages which introduce our interest to the GALT enzyme, the search for mutations allows to reach the html pages generated to explain the structural and functional effects of each mutation. Figure 1 shows an example of these pages. The template used provides:
- general information about the mutation, as Pubmed links to relevant article(s), references to databases;
- link to download the model of the mutant protein;
- summary of the effects detected by the analysis;
- details for each structural/functional feature analyzed.

We are aimed to apply the same procedure to investigate other proteins of interest for human pathologies.

ACKNOWLEDGEMENTS

This work has been partially supported by the Italia-USA programme “Farmacogenomica Oncologica”, and by Flag Project “Interomics”.

REFERENCES

Quantum.Ligand.Dock is an original modern method for in silico prediction of protein-ligand interactions via high performance docking code. It is the only docking method offering such a subtle supplement to protein docking algorithms as quantum entanglement contribution. The motivation hinges upon two arguments - the fundamental importance of quantum entanglement in molecular interactions and the realistic possibility to implement it by the availability of supercomputing power.

INTRODUCTION

Quantum Entanglement (1, 2) is emerging as a major topic in modern molecule science but still ignored in the analysis of recognition and current predictive docking approaches. In spite of the progress in prediction via in silico methods, intricacies in protein-ligand interactions are still beyond our reach. The crucial step is to focus on the precise description of the physics of protein-ligand interactions. The most reliable description is via ab initio quantum mechanical methods and the recent possibilities to access adequate computing power obliges the community to address the problem in the context of practical protein-ligand analysis tools.

Prediction of protein-protein and protein-ligand interactions via docking methods is at the focus of intense research. However all these methods do not face two issues – quantum effects and the self-consistency of electrostatic interactions (including the mutual influence of docking partners on their protonation states through interdependent perturbation of pKa values). Our contribution is the implementation of this essential but missing link in the context of protein-ligand interactions and its realization on a massively parallel GPU supercomputer via C / C++ / OpenCL programming environment. Thus we have developed ultrafast docking code with a strong potential for large scale systems biology projects. Concurrently we have put on a sound theoretical basis the interdependency of protein-ligand electric fields, the mutual influence on pKa values (ionization states) upon molecule encounter and the fundamentally important quantum entanglement effect. On the docking algorithmic side we make use of the significant speed-up of the Fast Fourier Transform parallelized effectively under OpenCL environment (3, 4). On the electrostatics side we apply an improvement of our own self-consistent and rigorous method GPU.proton.DOCK / PHEPS / PHEMTO (3-6), including modern Fast Multipole Methods in parallel version.

Side by side with these improvements we implement reasonable and practical approach for estimation of a fundamental quantum mechanism (quantum entanglement) which is emerging as a major topic in modern molecule science but still ignored in current docking approaches. Quantum entanglement contribution to the protein-ligand interaction is estimated via calculation of entangled states of the composite protein-ligand Hilbert space.

FIGURE 1. Quantum Entanglement in Protein-Ligand Recognition

Technically speaking, it is the tensor product $H_{\text{Protein}} \otimes H_{\text{Ligand}}$ of the Hilbert spaces of the protein molecule ($H_{\text{Protein}}$) and the ligand molecule ($H_{\text{Ligand}}$). For example using basis vectors $|0\rangle_{\text{Protein}}, |1\rangle_{\text{Protein}}$ for the Hilbert space $H_{\text{Protein}}$ and basis vectors $|0\rangle_{\text{Ligand}}, |1\rangle_{\text{Ligand}}$ for the Hilbert space $H_{\text{Ligand}}$ we can define the following entangled Bell state:

$$|\Psi\rangle = \frac{1}{\sqrt{2}} \left( |0\rangle_{\text{Protein}} \otimes |1\rangle_{\text{Ligand}} - |1\rangle_{\text{Protein}} \otimes |0\rangle_{\text{Ligand}} \right)$$

These calculations have motivic kinship to other issues such as the widely discussed and exciting quantum nonlocality in molecular systems including biological macromolecules.

REFERENCES

ULTRA RAPID, ACCURATE QUALITY ASSESSMENT OF PROTEIN STRUCTURE MODELS

Marcin J. Skwark1,2 & Arne Elofsson1,2*
Dept. of Biochemistry and Biophysics¹, Stockholm University,
Science for Life Laboratory², Stockholm, Sweden
arine@bioinfo.se

INTRODUCTION
Prediction of 3D structure of proteins is one of the major goals of contemporary bioinformatics. For each predicted model, there is a need for use of an independent measure to evaluate correctness of the model. This is the role of Model Quality Assessment programs (MQAPs). Traditionally, MQAPs focused on evaluating structural features of predicted models, to assess the likelihood of model being similar to a native protein structure. These approaches are capable of detecting non-physical model conformations, but of discriminating which of the two biophysically feasible structure models is in the correct conformation. The advent of consensus methods alleviated this problem. Consensus methods are based on the premise, that among different models of the same protein, the one that is most similar to the others is most likely to be correct.

Most of consensus-based MQAPs rely on structural superposition, which is a computationally expensive process and as such makes consensus approaches unfeasible for larger model ensembles. Additionally, structural superposition does not account for conformational flexibility of proteins.

METHOD
Approach presented in this work does not rely on structural superposition, but rather on comparison of inter-atom distance matrices. It is at least as efficient in selecting the most accurate models from the model ensemble, as world-leading consensus methods. The increase in selection accuracy is particularly notable in case of more difficult target, where there is no evident largest cluster of structurally similar models. Due to use of the streaming computing platform (off-the-shelf CUDA-compatible GPU), it is able to obtain at least a 10-fold speed-up in comparison to the other approaches, with no upper bounds on the amount of models in the ensemble, nor on the model size.

RESULTS
The method presented in this work – PconsD Q-score (GPU implementation of distance-driven quality metric) is significantly faster than other model quality assessment methods for non-trivial targets. It is up to 60 times faster than ModFOLDelustQ³ – another method relying on the same principle and approximately 8-10 times faster than superposition based methods, such as Pcons³. Additionally, it scales very well, both as far as the model length and model amount are concerned (see Figures 1 and 2).

Increased performance allows for much shorter turnaround times, thus enabling options not feasible with other approaches (e.g. iterative modelling, nearly real-time assessment etc.).

While, it is intuitively obvious that quality assessment methods based on distance matrix comparison do not correlate perfectly well with superposition based metrics, PconsD has been demonstrated to outperform superposition based methods as far as capability to select most native-like decoy in the set is concerned.

REFERENCES
A new de novo protein structure prediction method for transmembrane proteins (FILM3) is described, which is able to accurately predict the structures of large membrane proteins domains using a fragment selection approach in combination with a scoring function based solely on correlated mutations detected in multiple sequence alignments.

**INTRODUCTION**

Alpha-helical transmembrane proteins (TMPs) constitute roughly 30% of a typical genome and play critical roles in a diverse range of biological processes. Despite the recent increase in the number of solved TMP crystal structures, coverage of TMP fold space remains sparse. Computational methods to predict TMP structure are therefore vital in helping to further our knowledge of both the structure and function of these important proteins. Recently, we developed a new contact prediction approach called PSICOV (1), which makes use of sparse inverse covariance estimation techniques to very effectively overcome the indirect coupling effects that plague correlated mutation analysis of sequence alignments. Where sufficient numbers of homologous sequences are available, PSICOV prediction accuracy can approach 80%, even for long range contacts. To benefit from this information, we have modified the original FILM method (2) by replacing all knowledge-based potentials and other statistically derived scoring functions with a single scoring function based simply on contacts predicted by PSICOV.

**METHODS**

At the heart of FILM3 is an objective function that is entirely based on distance restraints that are inferred only from the multiple sequence alignment (MSA) and predicted transmembrane topology. PSICOV is first used to generate a list of predicted contacts from the target MSA. Where a contact is predicted, a constraint on the \( \text{C}^- \text{C}^- \text{C}^+ \) distance between the two given residues is applied. A further source of minimum distance constraints arises from our knowledge of the target’s predicted transmembrane topology and the simple meandering nature of alpha-helical transmembrane protein folds, allowing us to deduce approximate Z-coordinate values for residues in each transmembrane spanning segment, assuming a lipid bilayer thickness of 30Å. Model generation is then carried out in two phases: conformational searching and combinatorial refinement. Initial conformational searching is carried out using the standard FILM/FRAGFOLD approach (3), though with the standard energy function replaced by the distance constraint function described above. In addition, FILM3 uses Replica Exchange Monte Carlo to identify low energy conformations in place of simulated annealing. For each target, 100 independent runs were carried out each beginning with a randomly generated starting conformation, with and without Z-coordinate constraints. A combinatorial refinement step was then carried out using the final ensemble of models, in which the lowest energy model for the target was identified and the 100 lowest energy models fitted to it by rigid body superposition.
CHARACTERIZATION OF CATALYTIC SITE RESIDUES USING STING_DB STRUCTURAL DESCRIPTORS

Salim J. A. 1,2, Von Zuben F. J. 1, de Moraes F. R. 1,2, Neshich I. A. P. 1,2, Mazzoni I. 1, Jardine J. 1 & Neshich G. 1
Computational Biology, EmbraP and Dept. of Computer Engineering and Automation – UNICAMP 2.

The function and classification of a particular enzyme could in principle be obtained by selective description of its catalytic site residues (CSR). By using a set of STING_DB structural descriptors and humanly readable rules (enabling interpretation of results in terms of detailed characterization of nano environment for enzyme CSRs) we initiated construction of protein function “periodic table” based on CSR environment description.

INTRODUCTION
Chotia and Lesk 1 showed that protein function is maintained through conservation of its tridimensional structure more than of its primary sequence. There are several proteins with little or no sequence similarity and high structural similarity 2. The structural information is useful for explaining the biochemical mechanism of the protein function. However, there are evidences that proteins with only one fold possess several biological functions, while the same function could be seen for more than one fold (as in the TIM barrel fold (identified in enzymes classified by more than 60 different EC numbers)).

To lead with these challenges many simple descriptors have been used to improve the accuracy of methods that are designed to identify structural similarity, such as conservation descriptors. We shared with Bobadilla et al 3 the understanding of definition for specific active site as a physicochemical vicinity which provides for necessary nano environment conditions for protein functionality. Thus, a more generic characterization of the active site, based on structural descriptors (but without considering conversation properties) is proposed here in order to predict catalytic residues of enzymes and by doing so enable construction of classifiers for a protein function.

METHODS
Using the information provided by Catalytic Site Atlas – CSA 4 (only structures with literature supported entries) and PDBSprotEC 5 database, residues of enzyme structures were divided into two classes: the one containing catalytic site residues and another class with the remaining residues. For selected structures respective residue descriptors were extracted from STING 6 database.

Initial experiments using different classifiers, like decision trees and rule induction algorithms, have shown a limited performance in finding a humanly readable rule that could fully characterize the CSR nano environment and at the same time keep the catalytic residues grouped (classified). An interpretation of these preliminary results indicates that employed methods were retrieving information about specific residue properties rather than about shared environmental characteristics, mostly due to implementation of method’s local search and separate-and-conquer approaches.

Consequently, a global search method is needed to explore the environmental features of the catalytic site residues. Using a Genetic Algorithm with binary chromosomes of size equal to the number of descriptors (attributes) and a fitness function that executes an rule induction algorithm and evaluate its performance based on the prediction accuracy, number of rules were found and number of CSR groups identified (dyads, triads, tetrads, etc.). The latter step is very important for searching for rules that can describe the environment shared by CSR’s.

RESULTS & CONCLUSIONS
The generated rules for each specific reaction were able to capture the CSR of a majority of the structures in test set, including enzymes with different primary sequence and folding. The results for three EC numbers is showed in Table 1, where the rules found and the performance measures are presented for each EC number (the number of CSRs given in parenthesis). Identified rules work only for the same function enzymes. The structural alignment of Superoxide dismutase (EC: 1.15.1.1) and RNA polymerase (EC: 2.7.7.6) with two different folds are shown (Fig 1) - residues returned by the rule as CSR are in black.

The CSR of many different folds covered by the identified rules suggest that there is a conserved physicochemical environment shared among enzymes which catalyze the same reaction. Based on the interpretation of the rules it is possible to understand the mechanism of catalysis and also aid designing of new drugs.

TABLE 1 More details about the descriptors can be found at: http://www.ncbi.nlm.nih.gov/Structure/STING/megaHelp/MegaHelp_{PDB}.html#parameters

<table>
<thead>
<tr>
<th>EC</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.15.1.1</td>
<td>ch_repetitive_ene49 &amp; sponge_CA5.65 &amp; ep_AT_CA2.92 &amp; ced_CA3.5.0.21.7.0.066</td>
</tr>
<tr>
<td>3.4.2.1</td>
<td>ced_LHA_4,0.58 &amp; cpo_CA4 &amp; hydroRels-0.01 &amp; cpo_lha3 &amp; bhm_m_ene3.6 &amp; ch22-167.3</td>
</tr>
<tr>
<td>2.5.1.8</td>
<td>spenger_LHA3.2 &amp; cpo_CA2 &amp; density_CA4.4,0.79 &amp; density_CA6.6,0.82</td>
</tr>
<tr>
<td>3.6.1.16</td>
<td>167.3</td>
</tr>
</tbody>
</table>

TP/FP Rate (%) 96.4 / 0.02 70.4 / 0.01 85.4 / 0.02

REFERENCES
FRAGMENT-BASED LOOP MODELLING FOR X-RAY CRYSTALLOGRAPHY OF MEMBRANE PROTEINS

Sebastian Kelm1,*, Charlotte M. Deane1, Frank von Delft2

Dept. of Statistics1 and Structural Genomics Consortium2, University of Oxford. *kelm@stats.ox.ac.uk

Membrane proteins are hard to purify and crystallise and thus their structure prediction is an important focus of structural bioinformatics. Even after a membrane protein has been successfully crystallised, the data from the X-ray diffraction experiment needs to be interpreted. This process involves automatic three-dimensional modelling tools, which tend to fail when given low-resolution data such as is typically produced from membrane proteins. In an effort to improve model completion for X-ray crystallography we apply our fragment-based loop modelling method FREAD to soluble and membrane proteins at various resolutions and stages during the modelling and refinement process and combine it with crystallographic scores to explore new efficient model completion strategies.

INTRODUCTION

Membrane proteins are important drug targets but knowledge of their exact structure is limited to relatively few examples. Recent developments have lead to membrane-protein-specific 3D modelling tools, such as MEDELLER3, which achieve improved modelling accuracy over conventional methods designed for soluble proteins. In its original implementation, MEDELLER used the successful fragment-based loop modelling method FREAD3. Even though FREAD was not optimised for membrane proteins, it achieved higher accuracies than Modeller. We later demonstrated1 that, to achieve consistently good loop modelling accuracies, one should use a fragment database composed solely of membrane protein fragments. However, even highly accurate prediction tools rely on the availability of an experimentally determined template structure.

To speed up the experimental determination of membrane (and soluble) proteins, we now apply the FREAD algorithm to the problem of model completion in X-ray crystallography. Crystallographers rely more and more on automated methods for building and refining models that agree with the diffraction data. These automatic tools, however, perform poorly on low-resolution data such as that commonly obtained from membrane proteins and other difficult crystallography targets. Autobuilding programs generally produce incomplete and inaccurate models that the scientist then needs to manually correct and complete. This process is tedious and difficult. Furthermore, crystallographic software has to date made limited use of the state-of-the-art loop modelling techniques developed in the structure prediction community, where no electron density information is available. Combining such tools with even low-quality diffraction data should allow for large improvements in model quality.

In 2009, DiMaio et al. demonstrated4 that computational prediction tools such as Rosetta can be successfully combined with crystallographic approaches for molecular replacement, auto-building and refinement to enable the solution of structures whose diffraction data could previously not be interpreted. Our goal is thus two-fold: a) to develop a fast auto-building method for the more reliable parts of the model, which combines the knowledge encoded in a fragment database with X-ray diffraction data; b) to provide a semi-automatic method rendering the manual building of loops by crystallographers a simple and effective process.

METHODS

FREAD employs a database of protein fragments. The input to the program is a protein structure containing a gap, where 3D co-ordinates are missing. Only the amino acid sequence of this “loop” is known. FREAD uses this loop sequence, along with the structure of the adjacent “anchor” residues, to query its fragment database.

We combine FREAD with various crystallographic scores assessing map quality, the real-space fit of the model to the electron density map and agreements between the calculated and observed structure factors in reciprocal space.

Scoring schemes and auto-building strategies are explored by analyzing a set of crystallographic models at various stages before their final deposition in the PDB. Proteins are re-modelled from their initial molecular replacement solution using various combinations of loop modelling, real and reciprocal space refinement and model trimming. Accuracy is assessed on a separate test set of soluble and membrane protein structures, with the final deposited structure taken as the “correct” solution.

RESULTS & CONCLUSIONS

We show that combining the knowledge encoded in the FREAD database with the experimentally determined diffraction information allows for effective automatic model completion. We also provide crystallographers with a convenient tool for semi-automatic model building in the framework of the popular COOT suite.
**INTRODUCTION**

The evolution of proteins is the single process that has delivered the diversity and complexity of life that we see around us today.

While it is relatively simple to observe evolutionary change from sequence comparisons, it is a lot harder to discuss evolutionary transitions in structure space.

Since selective pressure for genetic change derives from improved, diversified or unaltered function [1], protein structure, as intimately linked to a protein's function is a vital unit in any consideration of the evolution of proteins [2].

Phylogenetic profiling, using genome annotations of structural families of proteins, can suggest an estimate for the last common ancestor of a fold: a coarse-grained phylogenetic ‘age’ of the structural family [3].

**METHODS**

Here we suggest various phylogenetic profiles of SCOP superfamilies using the predicted occurrences of structural domains on fully sequenced genomes, following a previously published method [3] (Figure 1). We obtain phylogenetic age estimates for 1,892 superfamilies.

**RESULTS & CONCLUSIONS**

We verify previous results showing a significantly higher age distribution for \(\alpha/\beta\) superfamilies than other classes. We further find many of the characteristic traits of the \(\alpha/\beta\) class, such as a parallel beta strand topology and an increased protein length positively correlate with age even within the different classes.

We explore the most common anti-parallel beta sheet motifs: the meander, greek key and jelly roll and find superfamilies containing a jelly roll are significantly younger than those containing just a greek key motif.

We also present results on a range of structural and sequence-based properties hypothesized to relate to the stability of a protein structure and show that for many of these properties, such as structural compactness and a hydrophobic core, the age distributions of superfamilies exhibiting signs for either extremely high or extremely low thermostability are significantly lower than those with more moderate values.

**REFERENCES**

TARGETING DRUG-RESISTANT HIV-1 PROTEASE MUTANTS
USING INHERENT PROTEIN FLEXIBILITY

Yang Shen1,2* & Bruce Tidor1,3

Depts. of Biological Engineering and Electrical Engineering & Computer Science, Massachusetts Institute of Technology1 and Toyota Technological Institute at Chicago2. *Tidor@mit.edu or YangShen@ttic.edu

Drug-resistant target mutation has been one of the major obstacles in delivering effective therapeutics. When limited information of drug-resistant mutants is available, can target flexibility be incorporated to design low-susceptibility drugs? We address this question using the prototype example of HIV-1 protease inhibition. An inhibitor library was computationally designed to target an ensemble of simulation structures from molecular dynamics (MD) and also an ensemble of wild-type and drug-resistant mutant X-ray crystal structures. A susceptibility classifier was trained from the dynamics ensemble and achieved high accuracy and hit rate for the mutant ensemble. The results suggest that inherent protein flexibility demands similar broadly-binding inhibitors as mutants do. Properties in the classifier also provide useful insights into underlying molecular mechanisms.

INTRODUCTION
It is challenging to apply structure-based drug design to “moving” targets that quickly acquire drug-resistance. Explicit consideration of all forms can be daunting if not impossible. One may overcome this problem with better understanding of available structures. By analyzing a set of X-ray structures for HIV-1 protease, Zoete et al. discovered that their structural variances reflected inherent protein flexibility regardless of sequence or ligand specific.

Inhibited by their findings, here we introduce an approach to prepare inhibitors for drug-resistant mutants by targeting an ensemble reflecting protein flexibility. Results in the prototype case of HIV-1 protease inhibition suggest its potential for developing design principles and uncovering molecular mechanisms for drugs that target mutations with limited sequence or structure information.

METHODS

Target Preparation. A 20-ns MD simulation was run for a wild-type protease complex (PDB code 2Q16) with explicit solvent. The 14 middle snapshots sampled at every 1 ns were retained as a training target ensemble to capture inherent protein flexibility. In addition, a test target ensemble was collected to include 4 other wild-type and 10 representative drug-resistant mutant structures.

Computational Inhibitor Design. All inhibitors consisted of a scaffold derived from darunavir plus three functional groups selected from the ZINC database and commercial catalogs. Each design was performed in a hierarchical way. A pruning, enumerating, and re-scoring approach produced an ordered list of designs.2

Classifier Training. Each inhibitor was characterized with 36 molecular descriptors calculated from QikProp and assigned a susceptibility score (+1 if binding to more than 50% of the training or the test ensemble, and –1 otherwise). A linear classifier was trained with an L1-regularized logistic regression implemented in the LIBLINEAR package.

RESULTS & CONCLUSIONS

12728 (17906) inhibitors were designed to target the training (test) target ensemble, with 3070 shared by both. 42 (165) inhibitors were classified as being of low susceptibility to target change. They were outnumbered by more than 100-fold by undesired, highly susceptible ones, which highlights the challenge of designing and screening them.

Using inhibitors targeting the training ensemble reflecting inherent protein flexibility, a series of linear classifiers was constructed by balancing classification error and model complexity. Only 10 of the 36 descriptors were needed to achieve over 80% in balanced accuracy for the test drug-resistant ensemble. A measure relevant to hit-to-lead screening, the positive predictive rate, was 3–5 fold greater.

Linearity was exploited for interpretability, with greatest weights assigned for molecular size, flexibility, and hydrophobicity. For each descriptor, low-susceptibility inhibitors for either set tended to be in a narrow range compared to highly susceptible ones. They also tended to be smaller, more flexible, and less polar (Figure 1).

These results suggested that protein flexibility posed similar physicochemical demands for broadly-binding, low-susceptibility inhibitors as did the set of drug-resistant mutants. This finding, a dual form of what Zoete et al. found, provides a promising drug-discovery approach to target many uncharacterized mutants with inherent flexibility modeled from a few available structures.

ACKNOWLEDGEMENT

We thank Bracken King for providing molecular dynamics ensemble. This work was supported by the National Institutes of Health (GM065418 and GM082209 to BT) and Toyota Technological Institute (research funds to YS).

REFERENCES

INTRODUCTION

The estrogen receptor is a biologically important protein with crucial roles both in normal physiology and in breast cancer. We are using molecular dynamics simulations to study the ligand-induced conformational changes in the ligand-binding domain (LBD) of the estrogen receptor-alpha. We use a measure we developed previously, mean-contact-deviation (MCD), to compare the dynamics of the 1QKU estradiol-bound LBD structure (with estradiol removed) to a ligand-free homology model based on the 1LBD RXR-alpha structure. MCD generalizes the widely used root-mean-squared-deviation (RMSD) measure from three dimensions to n-dimensions, where n in the current study equals the number of atoms either in the ligand-binding pocket or in the complete monomer. We compare individual atom trajectories for both the RMSD and the MCD analyses.

METHODS

In previous work, we used a piece-wise linear cut-off function to define the contact between two atoms. When the cut-off used was too large, this resulted in contact maps that were not positive-definite. In this work, we used a Gaussian function rather than a discrete cut-off value, with the parameter σ defining the inflection point. We define the “smoothed” contact between atoms i and j to be

\[ c_{ij} = e^{-\frac{d_{ij}^2}{\sigma^2}} \]

where \( d_{ij} \) is the distance between atoms i and j. For sufficiently small \( \sigma \), there exists a positive-definite, smooth contact matrix \( C = (c_{ij}) \) which can be used to define the generalize dot product \( \langle u, v \rangle_C = u^T C v \). Associate each atom i with the standard unit vector \( e_i = (0, ..., 1, 0, ..., 0) \). Observe that \( \langle e_i, e_j \rangle_C = c_{ij} \). There exists a linear transformation \( T = R e_i \), where \( R = V D V^T \), which transforms the above generalised dot product into the standard one. Specifically,

\[ \langle r_i, r_j \rangle = \langle e_i, e_j \rangle_C = c_{ij} \]

The vector \( r_i \) is the intrinsic contact vector of atom i. The relative MCD of atom i for N frames of a molecular dynamics simulation is defined to equal:

\[ \frac{1}{N} \sum_{i=1}^{N} \left[ 1 - r_i^T (k) P_i (k + 1) \right] \]

where \( r_i \) is the vector obtained by superimposing, in n-dimensional space, the intrinsic contact vectors of frame \( k+1 \) on the intrinsic contact vectors of frame \( k \).

RESULTS & CONCLUSIONS

While relative atom RMSD is a measure of structural deformation, relative atom MCD behaves more like an atom “chaffing” index. Comparisons based on MCD indicate that the ligand-binding pocket atoms exhibit more “chaffing” in the model than in the 1QKU structure. In contrast, RMSD did not indicate any differences in structural deformation of the ligand-binding pocket. When analyzing all atoms in the monomer, MCD highlighted histidine 524, a residue experimentally observed to be important for ligand binding, as especially flexible in the model. In contrast, histidine 524 did not exhibit significant differences in RMSD. However, RMSD analysis suggested significant structural differences in the N-terminal end of helix 3, while MCD did not. (See Figure 2.) We believe, therefore, that RMSD and MCD detect different functionally relevant changes, and it is likely that the two analysis techniques offer complementary indications of protein structural changes.

The code (written in Octave) used to compute the results in this paper is available at www.rose-hulman.edu/~shibberu/ContactGeometry.
We recently solved the structure-matching problem for protein oligomers and large molecular complexes, resulting in an ultra-fast structure search engine and tools for matching oligomer structures containing well over 10,000 amino acid residues\textsuperscript{1,2}. With these tools it is possible to scan the complete database of biologically active units of protein structures (~80,000 assemblies) within seconds. Here, we use these techniques for the comparative analysis of large multiple-chain assemblies of TIM-barrels and report hitherto unseen structural correlations.

**METHODS**

We used the TopSearch\textsuperscript{3} web service to scan all ~80,000 publicly available biological assemblies of proteins for structural similarities to a recently determined dodecameric form of Pdx1 enzyme from *Plasmodium berghei* (PDB code: 4adt, Figure 1). Subsequent pairwise structure comparison, visualization and detailed investigation of the structural matches found were carried out with the TopMatch web service\textsuperscript{1}.

**RESULTS & CONCLUSIONS**

We find significant structural similarity to Pdx1 for a variety of homomeric assemblies of TIM-barrel chains. These similarities frequently exceed the similarities expected from a pairwise match of single TIM-barrels and are due to the preservation of the relative arrangement and orientation of the TIM-barrel monomers. Surprisingly, such recurrent arrangements are also observed in proteins of otherwise highly different quaternary structure (Figure 2). In view of the diverse set of oligomeric interfaces found in these proteins and the low sequence similarity between the matched parts, structural correlations of this kind point to intriguing yet unknown principles of protein structure assembly.

**REFERENCES**

INTRODUCTION
Cholera, diphtheria and pertussis toxins are well-known examples of a family of potent toxins that covalently modify intracellular proteins in target cells. The toxin ADPRTs can be divided into three structural types: the CT-PT subgroup (CATH: 3.90.210.10), the C2/C3 subgroup (CATH: 3.90.176.10) and the DT group (CATH: 3.90.175.10). One feature of this family is low primary sequence identity, but high structural similarity. Thus, here we have employed an exhaustive fold-recognition search strategy to identify new toxins from both the CT-PT and C2/C3 subgroups. This work is complemented with structural modeling that uses spatial restraints from homologous structures and physics-based refinement. Finally, we present cell-based model validation. We also discuss continuing inhibitor development initiatives.

METHODS
(1) Remote homolog detection. Due to low sequence identity, candidate toxins were identified using fold recognition databases in conjunction with the sensitive profile-profile search method, HHsenser. The candidate sequences were filtered (keeping only sequences from bacterial pathogens that are predicted to be secreted, and ensuring the ADPRT regular expression pattern was present). (2) Structural modeling. We undertook template-based modeling for several of the new toxins, using both MODELLER (to capture evolutionary information) and also the Molecular Operating Environment (MOE) and Rosetta (for physics-based refinement, for example, a physically-realistic energy function). Models were analyzed using MOE. (3) Cell-based validation. A cell-based assay in which the candidate toxin was expressed in yeast under the control of a copper-inducible promoter was used to confirm activity.

RESULTS & CONCLUSIONS
Several of the newly identified toxins from this project were outlined in-depth previously. In addition to these toxins, we have identified C2/C3-like toxins in Pseudomonas fluorescens (e.g., Flare), and CT-PT-like toxins in Citrobacter rodentium (e.g., Citrin) and Paeonibacillus larvae (e.g., Larvin2), among others. These three toxins have been confirmed active in a yeast cell-based assay. Toxin expressed in yeast under the control of a copper-inducible promoter resulted in a growth-defective phenotype. Furthermore, we have conducted active-site structural modeling to both satisfy spatial restraints from homology and to take advantage of a modern, all-atom, high-resolution energy function. We identified likely toxin-NAD\(^+\) interactions (at sites critically important for binding/catalysis in ADPRTs) and confirmed these using alanine substitution that restored the growth phenotype in yeast (Citrin Q138A, E140A, Q138A/E140A; Larvin2 Q155A, E157A and Q155A/E157A; and Flare E1558A, E1560A and E1558A/E1560A). These interactions partially validated the models, and may also serve as the basis for continued inhibitor development.

REFERENCES
INTRODUCTION

Human African trypanosomiasis (aka “Sleeping Sickness”) is caused by the protozoan parasite Trypanosoma brucei (T. brucei). New treatments are urgently needed, and proteases are some possible drug targets. Proteases catalyze the hydrolytic breakdown of proteins, and the full complement of an organism’s proteases is termed its “degradome.” In drug design, surveying homologs of target proteins can reveal unanticipated interactions [1], and, conversely, some drugs may show efficacy with unanticipated targets making them useful in treating diseases other than those for which they were designed. Broad comparisons of homologs are now possible using public databases of protein sequences and structures. We constructed the first global views of sequence and structure similarity, incorporating crystal structures and models, of the predicted active proteases of T. brucei and its human host, allowing side-by-side comparisons. We aimed to see how comparisons of the two degradomes might lead to new insights for drug design. We include a detailed structural evaluation of two groups of parasite proteases that may have potential as new drug targets.

METHODS

Predicted proteases were identified from the T. brucei genome [2] by a BLAST [3] search of MEROPS [4], and false positives removed by analyses using Swiss-Prot and Pfam. Human protease sequences were from MEROPS and the Mammalian Degradome Database [5]. Active proteases were identified using tools at the MEROPS website. Crystal structures were from the PDB, and models were from MODBASE [6] and Prime (Portland OR). Sequence and structure similarity networks using all-by-all BLAST and FAST [7] scores, respectively, were visualized with Cytoscape [8]. Recombinant TbM32 was cloned, expressed and purified using standard molecular biology techniques, and assayed using the synthetic carboxypeptidase substrate FAFF. Inhibitors were 1.10P, lisinopril and ACE2 inhibitor 28FII, a gift from the Dives laboratory [9].

RESULTS & CONCLUSIONS

Two T. brucei-specific clusters were identified that were far in sequence space from human proteases. The first, the cysteine protease C51 family, was also quite different in 3-D structure from human proteases, suggesting that C51 proteases may be good drug targets. The second, “TbM32,” a singleton of the metallocarboxypeptidase M32 family, had several close human structural homologs. Although structural similarity of a trypanosome M32 protease to human ACE has been previously recognized, functional and active site similarities of TbM32 to human ACE2 indicates that inhibitors for ACE2 may be a better starting point for anti-parasite inhibitor. Our assays confirmed that TbM32 is a carboxypeptidase, like ACE2 and unlike ACE (which is a dipeptidyl peptidase). The presence of an arginine (R273) in the ACE2 S1 binding pocket, instead of the corresponding Glu in ACE, creates a smaller pocket in ACE2 that allows only one residue instead of two to fit into the active site C-terminal to the cleavage point. This difference also helps to rationalize the high selectivity of inhibitors for ACE relative to ACE2 [10]. As shown in Figure 1, TbM32 has a corresponding arginine (R348) in this position. We then assayed the recombinant TbM32 with ACE2 inhibitor 28FII (10 µM). The ACE2 inhibitor produced a significant inhibition of TbM32 whereas the ACE inhibitor lisinopril did not.

![Figure 1. The structural alignment of the model of TbM32 (”,TbM32m,” purple) with ACE2 (PDB 1R4L, yellow) predicts a steric clash (arrow) of ACE2 R273 with ACE inhibitor lisinopril (orange) which was superimposed from ACE (1O36). ACE2 R273 and the corresponding R348 from TbM32m are predicted not to clash with an ACE2 inhibitor (green, MLN4760 from 1R4L).](resources.html)
ALLOSTERIC COMMUNICATIONS IN THE OUTER-MEMBRANE TRANSLLOCATION DOMAIN OF THE PapC USHER PROTEIN?

Irene Farabella¹, Gabriel Waksman and Maya Topf”

¹Institute of Structural and Molecular Biology, Birkbeck College and University College London, University of London, London WC1E 7HX.

Corresponding Author: Maya Topf, m.topf@cryst.bbk.ac.uk.

Combining the information from prolonged molecular dynamics simulations on the native PapC translocation domain in a POPE-POPG lipid bilayer as well as different mutants lacking an α-helix, a β-hairpin, and both and the sequence conservation and mutual information-based co-evolution analyses of PapC ushers, we identified important residues that mediate the allosteric signal transmission between these elements and the plug domain in the translocation domain of PapC usher.

INTRODUCTION

The PapC usher protein is the outer membrane platform in the chaperone-usher pathway responsible for the assembly and secretion of P pilus. This adhesive appendage is responsible for host attachment and recognition in uropathogenic Escherichia coli (UEPC). The usher comprises a transmembrane domain (TD) composed by a translocation pore (TP) interrupted by a conserved plug domain (PD), and an N- and two C-terminal periplasmic domains [1]. The structure of the apo-TD consists of a 24 β-stranded barrel (TP) where the PD is inserted into the loop connecting two β strands (β6 and β7) via two long linkers (p-linker1 and p-linker2), occluding the luminal volume of the pore [2,3]. In the usher-chaperone-adesin complex [1] the PD is displaced to accommodate the adhesin. Studies suggest that this gating mechanism is likely controlled by two unique secondary structures elements of the TP: a β-hairpin and an α-helix [3,4]. Here we investigate the role of these two secondary structure elements in the context of the allosteric signal propagation mechanism within the TD by integrating all-atoms molecular dynamics simulations (AA-MD) and sequence evolution information.

METHODS

Model systems of the native PapC-TD (sim1) and of 3 mutants lacking: (i) β-hairpin (sim2), (ii) the α-helix (sim3), (iii) and both (sim4), were generated using MODELLER-9v7 [5] using the PapC-TD X-ray structure (PDB ID: 2vqi [3]) as template. Each of the systems was embedded in a POPE/POPG (3:1) lipid bilayer and solvated in a 0.15 M NaCl solution. The systems were then energy minimized and equilibrated in a multisite process. For each system AA-MD (∼70-72 ns of unrestrained simulation) was performed with the Gromacs package [6] using the OPLSA-AA force-field (ff) for the protein and the Berger ff for the lipids. The occupancy of the hydrogen bonds and salt bridge for each system was identified and a non-covalent interaction matrix was build using the normalized occupancy (interaction occupancy). A structure-based multiple sequence alignment (MSA) of the PapC-TD homolog was obtained using Expresso (3DCoffee) [7]. Conservation and co-evolution information from the MSA were obtained with: (i) Consurf web server (conservation) [8] and (ii) Normalized Mutual Information (NMI) (co-evolution) [9].

RESULTS & CONCLUSIONS

Based on the Root Mean Square Fluctuation (RMSF) analysis of the four MD trajectories, the native TD and the mutant systems show only modest differences in fluctuations. Interestingly in the absence of the α-helix and of the β-hairpin (sim4) a significant increase in fluctuations occurs in a distal region of the protein (p-linker1, β8-β9oop and β12-β13 loop). To understand the communication within this region we analysed the non-covalent interaction occupancy in the different systems. The native TD is characterized by the existence of extensive strong non-covalent networks within the barrel strands as well as strong interactions between the TP and the α-helix and the β-hairpin. The non-covalent interactions between the PD and the TP are mostly weak, although the absence of the β-hairpin (sim2) seems to affect a few strong interactions. The TP is communicating with p-linker1 and p-linker2 via a network of highly stable non-covalent interactions that show significant changes (Δoccupancy>2σ) in the mutants systems, especially in the absence of the helix (sim3-4). Interestingly, some of the residues that were identified to play a crucial role in the average non-covalent analysis appear to be evolutionary important residues. With the integration of these two types of analysis we identified a set of important residues on: (i) p-linker1 (D104, Y115), (ii) p-linker2 (T186, Y184, A180, V182, S181), (iii) β12 (R277), (iv) β13 (R297, F293, I291), (v) β7 (R192) and (vi) β8 (S218) that are important for the propagation of the allosteric signal within the TD.

REFERENCES

EVOLUTION OF PROTEINS STUDIED BY SECONDARY STRUCTURE COMPOSITION

J. Dongun Kim1,2, Stefan Senn1, Arik Harel1, David A. Case0, Vikas Nanda1, Paul G. Falkowski1,2
Department of Chemistry and Chemical Biology, Rutgers University1, Institute of Marine and Coastal Sciences, Rutgers University1, BioMaPS Institute for Quantitative Biology, Rutgers University1, Department of Biochemistry, University of Medicine and Dentistry of New Jersey2. jk@marine.rutgers.edu

We aim to construct the deep evolutionary history of the metal containing enzymes responsible for the critical electron transfer reactions across the tree of life. Using structural information, we are developing independent phylogenetic trees of how biologically catalyzed electron transfer reactions (the microbial “engines”) evolved. Herein, we report a structure profiling approach for comparing evolutionary distances across metal catalytic sites.

INTRODUCTION

We developed a method to calculate similarity based on protein secondary structure compositions near and around catalytic metal sites. The underlying hypothesis is an extension from our previous work that the protein secondary structure around the metal center would be more conserved, as it should obey the rules set by metal coordination chemistry. To our best knowledge, this is the first attempt to infer phylogenetic distances between distant protein folds based secondary structure compositions (“composition profiling” method).

METHODS

We first defined the secondary structure based on DSSP, a database for protein secondary structure, and developed a scoring function based on amino acid residue to catalytic site distance. Protein structures were sorted according to gold standard domains determined from other work recently published by our group2. Compositions of secondary structures were typically very similar for structures within the same domains, thus the average composome (compositional profile) was obtained for representing a given gold standard domain. Using the average composome for each gold standard domain, a distance matrix was generated. Fitch-Margoliash algorithm with global tree optimization was used for the construction of the preliminary phylogenetic tree of protein domains (Figure 1).

RESULTS & CONCLUSIONS

The compositional finger printing method for calculating evolutionary distances across folds collapses the rich 3-dimensional information into a single distance. To evaluate the validity of coarse-grained composome method, or reduction in dimensionality, we will compare distances to those obtained from 3D structural superimposition of both metal binding pockets and whole structures. Structural superimposition method alone is also very challenging because (a) pair-wise comparison requires high computational effort and extensive manual assessment of the results and (b) fold relationships can be lost across wide range of fold complexity due to the high precision of the superimposition method. Once we validate the composome method, general evolutionary constraints will be inferred based on secondary structure statistics, while 3D-structural superimposition method will calculate a precise pair-wise evolutionary distance based on structure. Both methods in parallel will synergistically reveal the history of protein folds.

REFERENCES

RETHINKING MOLECULAR SIMILARITY: COMPARING COMPOUNDS BASED ON BIOACTIVITY

Paula M. Petrone
Chemoinformatics and Statistics, Hoffman – La-Roche. paula.petrone@roche.com

Since the advent of High Throughput Screening (HTS), there has been an urgent need for methods that facilitate the interrogation of large-scale chemical biology data to build a mode of action (MoA) hypothesis. We develop a tool that compares compounds solely based on their bioactivity -the chemical biological descriptors “High-Throughput Screening Fingerprints” (HTS-FP). In the current embodiment, data are aggregated from 195 biochemical and cell-based assays developed at Novartis and can be used to identify bioactivity relationships among the in-house collection comprising ~1.5 million compounds. We demonstrate the value of the HTS-FP for virtual screening and in particular scaffold hopping outperforming state of the art methods in several aspects, retrieving bioactive compounds with remarkable chemical dissimilarity to a probe structure.

INTRODUCTION

With the advent of high throughput screening (HTS), relationships between compounds and biological entities have been studied on an enormous scale. Typically, MoA hypotheses are based on the assumption that structurally chemical compounds are likely to share similar properties and will bind to the same group of proteins. Chemometric approaches that rely on the use chemical descriptors to build quantitative structure-activity relationships (QSAR) have been geared towards predicting activity against a target. One reason why these models often do not live up to expectations is the rugged and high dimensional nature of the activity landscape. By construction chemical similarity cannot explain the activity of a compound against a specific pathway or groups of pathways which may or may not be known.

In our work, we investigate the applicability of bioactivity comparisons between compounds as a means for virtual screening and library design on an unprecedented scale. We have developed a set of biological descriptors, termed “High-Throughput Screening Fingerprint” (HTS-FP), which translate the wealth of HTS data into a form that can be readily interrogated by computational methods. When applied to virtual screening, HTS-FP has the capability of discovering novel active chemotypes for a phenotype (scaffold hopping) because it uses no structural information when comparing compounds. Also, HTS-FP can be employed to generate subsets of biologically active compounds, allowing for more efficient identification of active compounds when entire libraries cannot be screened.

METHODS

We use data from 195 assays developed at Novartis over a timeframe of 10 years, which cover a broad variety of protein families and technologies including fluorescence intensity, radioactivity and mass spectrometry. By comparing the similarity of compounds based on their HTS-FP, we elucidate bioactivity relationships among the in-house collection of ~1.5 million compounds. We have incorporated into our fingerprint both biochemical and cell-based assays. The advantage of using biochemical assay data is that it adds target-specificity to the fingerprint. This is especially powerful in cases were traditional descriptors would not perform well: the case for two structurally different compounds which inhibit the same enzyme by binding to different pockets. The advantage of utilizing cell-based assays is that many of these assays target an entire functional pathway or high-level phenotype, rather than the ability of a molecule to bind to a specific protein. Comparing compounds’ cell-based activity profiles can lead to the identification of compounds which produce a similar phenotype, yet not necessarily operate through the same mode of action. Conversely, comparing activity profiles of independent biochemical assays could not easily lead to such associations.

RESULTS & CONCLUSIONS

We have shown that fingerprints based on biological activity profiles can provide effective predictions on the cellular response of compounds without recourse to chemical structure similarity. By introducing biodiversity in HTS libraries, we can increase not only the hit rate but also the chemical diversity in hit compounds identified. Previous studies by Keiser et al. show that protein targets may be quantitatively related by the chemical similarity of their ligands. Here we show that the reverse is also true: even in the absence of chemical similarity, ligands may be quantitatively grouped by the biological closeness of their targets. In this context, HTS-FP therefore become a powerful way of finding those compounds that are bioactively but not structurally related. HTS-FP present an alternative approach to similarity searches, revealing additional chemotypes that open new opportunities in chemical and patent space.

ACKNOWLEDGMENTS

Benjamin Simms, Florian Nigsch, Eugen Loukine, Peter Kutchukian, Allen Gornett, Zhan Deng, John W. Davies, Jeremy Jenkins and Meir Glick at the Lead Discovery Center, Novartis Institutes for Biomedical Research. NIBR Education Office for financial support.

REFERENCES
EXPLORING THE EVOLUTION OF PROTEIN FUNCTION IN ARCHAEA

Alexander Goncearenco1,2,3 & Igor N Berezovsky1
Computational Biology Unit1 and Dept. of Informatics2, University of Bergen, Norway. 1ago064@uib.no

The evolution of enzymatic function did not start from complete protein domains. How did the original enzymatic domains emerge and what were their building blocks? Our analysis of the archaeal superkingdom yields the dominating mechanisms in different periods of protein evolution, which resulted in several levels of organization of biochemical function.

INTRODUCTION

Protein evolution and evolution of protein function, in particular, is a long-standing topic of keen interest in both experimental and theoretical aspects. The progress in experimental studies revealed the structure, biochemical function and catalytic mechanisms of many proteins. However, the questions how the first enzymes emerged and what their building blocks were remain unresolved.

The first task is to define the unit of enzymatic function. Enzymes are complex catalytic machines performing biochemical transformations as sequences of elementary chemical reactions. Therefore, the biochemical functions can be represented as sets of elementary ones. Previously, we explored how the polymer nature of proteins determined the units of protein structure and function and introduced the concept of elementary functional loops (EFLs)2. We hypothesized that EFLs, 25-30 residue long segments, are the functional units of enzymes, possessing distinct signatures and providing catalytic and binding amino acids to the enzyme's active sites. Some of the EFLs are presumably descendants of primordial catalytic peptides that preceded enzymes in the protein-RNA world. Though severely changed, structural and functional “signatures” of these ancestors could have survived in contemporary functional motifs.

METHODS

The functional signatures of elementary functional loops were reconstructed in form of sequence profiles. We used the functional annotation of archaeal clusters of orthologous groups (arCOGs)3, which splits the biochemical functions into the most ancient and common in the “core” and relatively new and specific functions in the “shell”.

RESULTS

We analyzed distant evolutionary connections between protein functions in Archaea based on the EFLs comprising them (Figure 1). One elementary function can work in different enzymes, forming their biochemical functions in combinations with other EFLs. Descendants of a particular ancestral elementary function can be found in unrelated folds and functions. Therefore, evolutionary connections unraveled by EFLs go beyond homology on the functional superfamilies level, illuminating the very process of building functional domains from the elementary units. We show examples of the EFLs in new functional domains, as well as reutilization of EFLs and functional domains in building multidomain structures and protein complexes.

CONCLUSIONS

Our analysis of the archaeal superkingdom yields the dominating mechanisms in different periods of protein evolution, which resulted in several levels of organization of biochemical function.

REFERENCES

PREDICTING RESIDUE-RESIDUE CONTACTS AND DISORDERED REGIONS WITH DEEP NETWORKS

Jesse Eickholt1,2,3 and Jianlin Cheng1,2,3,*

Dept. of Computer Science1, Informatics Institute1 & C. Bond Life Science1, University of Missouri.
*echj@missouri.edu

Recent developments in technology and methodology now allow for effective and efficient training of large, deep networks. These networks can be trained rapidly and assembled into boosted ensembles. Here we present a novel approach to predicting residue-residue contacts and disordered regions from sequence using deep networks.

INTRODUCTION

New advances in technology and statistical machine learning approaches now make it possible to train very large, deep neural networks. Using GPUs and CUDA, classifiers can be trained in a rapid fashion on very large datasets. Given the large datasets associated with protein residue-residue contacts and disorder and the difficulty of predicting contacts and disordered regions, we have developed new approaches to predict residue-residue contacts and disorder using deep networks and boosting.

METHODS

The basic framework used to predict residue-residue contacts and disorder regions was a combination of restricted Boltzmann machines (RBM) trained to form deep networks (DN). Training a DN involves learning layers of RBMs with contrastive divergence and then fine tuning the entire network with the back propagation algorithm12. The entire training procedure was implemented with matrix operations and utilizedCUDA3, a python library providing access to CUDA enhanced matrix calculations. Using CUDA and GPUs allowed us to fit large deep networks with over a 1 million parameters in under an hour.

To enhance the performance of the classifiers, we created boosted ensembles using a variant of AdaBoost4. In particular, when training a new classifier for the ensemble we sampled from a set of training examples. The probability of including a training example was adjusted according to the performance of previous classifiers. The final prediction was a performance weighted combination of the predictions from each DN in the ensemble.

For the prediction of medium and long range residue-residue contacts we used a number of structural features predicted from sequence, pair wise potentials and a few global features. For the prediction of residue and disorder using deep networks and boosting.

RESULTS & CONCLUSIONS

To determine the effectiveness of the deep network contact predictor we compared it with SVMcon1 on two CASP9 datasets. Table 1 shows the average accuracy and coverage of the top L and L/5 medium and long range contact predictions on a set of 16 targets comprised of hard domains (ie, FM and FM/TBM domains). Table 2 reports the performance of our method on 111 valid CASP9 targets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc(L)</th>
<th>Acc(L/5)</th>
<th>Med</th>
<th>Acc(L)</th>
<th>Acc(L/5)</th>
<th>Med</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVMcon</td>
<td>0.12</td>
<td>0.15</td>
<td>0.20</td>
<td>0.09</td>
<td>0.37</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 1. The accuracy and coverage of the top L and L/5 medium and long range predicted contacts on 16 hard targets. L is the length of the protein.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc(L)</th>
<th>Acc(L/5)</th>
<th>Med</th>
<th>Acc(L)</th>
<th>Acc(L/5)</th>
<th>Med</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVMcon</td>
<td>0.14</td>
<td>0.19</td>
<td>0.29</td>
<td>0.36</td>
<td>0.13</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Table 2. The accuracy and coverage of the top L and L/5 medium and long range predicted contacts on 111 CASP9 targets. L is the length of the protein.

To evaluate the deep network disorder predictor, we used 117 CASP9 targets and calculated the area under the ROC curve and balanced accuracy. For these targets, the area under the ROC curve was 0.82 and the method achieved a balanced accuracy of 0.74. This level of performance would have placed the method among the top sequence based residue disorder predictors in CASP96.

The work was partially supported by a NIH grant (5R01GM093123) to JC and a NLM fellowship to JE.

REFERENCES

MODELING RNA $\chi$ ANGLE FLEXIBILITY IN A SOFTWARE FOR STRUCTURE-BASED RATIONAL DESIGN OF RNA AND RNP COMPLEXES

Swati Jain$^{1,2}$, Laura Murray$^{3}$, Jane S Richardson$^{4}$, Bruce R Donald$^{4,*}$ & David C Richardson$^{3}$

Computational Biology and Bioinformatics Program$^{4}$, Dept of Biochemistry$^{4}$ and Computer Science$^{4}$, Duke University, Molecular Biophysics and Biochemistry$^{3}$, Yale University. *swati.jain@duke.edu

INTRODUCTION

RNA, a vital biological macromolecule, performs a variety of important cellular functions. RNA molecules participate in protein synthesis, regulate all aspects of gene expression, catalyze chemical reactions and control metabolic processes. For performing most of these functions, RNA binds to proteins to assemble into functional RibonucleoProtein (RNP) complexes. The timely and correct formation of these complexes is important for normal functioning of the cell. Therefore an emerging strategy for therapeutic intervention is to design novel molecules to inhibit the formation or block the function of these complexes.

RNA molecules are capable of folding into very intricate and complex three-dimensional structures. RNA structure is also very complex and flexible and changes conformations (involving both small scale base movements and large scale backbone movements) while binding proteins and small molecules. Unfortunately, the nature of these conformational changes is not clearly understood and characterized.

Computational structure-based rational design strategy is one of the techniques that can be used to design RNA and RNP complexes. However, the design software needs to take into account the structural accommodations RNA molecules can undergo in order to achieve good and realistic designs.

Here we describe the work we have done to understand different levels of RNA structural flexibility observed in high-quality RNA structures and modeling that flexibility into design algorithm that can be used to achieve realistic results.

METHODS

We have assembled a non-redundant and quality conscious dataset of RNA Crystal structures, called RNA11. This dataset contains high-quality structures all with resolution 3 A$^\circ$ or better. All PDB files containing RNA residues, released on the PDB website on or before 11/11/11 were considered as candidates to be a part of the dataset. The PDB files were analyzed on MolProbity, a structure validation server developed in our lab, and the structures were hand picked from the candidate structures depending on the quality of their validation statistics (which includes atom-atom clashes, geometry outliers, pucker outliers). The result is a dataset of 311 files, with close to 25,000 RNA residues. This dataset has been used to study RNA structure, its intrinsic features and flexibility, both in terms of backbone and base dihedral angles.

The Donald lab at Duke University has developed an ensemble-based protein design algorithm OSPREY, which uses provable and deterministic algorithms like Dead-End Elimination and A$^*$ search to guarantee to find the optimal solution for a given set of parameters. The lad has also extended DEE to incorporate protein side-chain and backbone flexibility in the design process. The software has been shown to work on a variety of design problems, including active-site redesign, predicting resistance mutations, and design of protein-protein interfaces. We have extended this used this software to design RNA and RNP complexes.

RESULTS & CONCLUSIONS

Our analysis on the RNA11 dataset shows significant differences in the range of values observed for the $\chi$ angle (the dihedral angle across the glycosidic bond) for C3'-endo versus the C2'-endo ribose pucker, as well as for purines (A, G) versus the pyrimidines (C, U). The values for purines populate three clusters in both C3'-endo and C2'-endo sugar puckerers, whereas pyrimidines populate one cluster. These values are now used for structure validation in Molprobity and structure refinement in PHENIX.

These range of values for $\chi$ angles have been used to model RNA base flexibility in the OSPREY design software. The range of values observed above are sampled are regular intervals and given as input to the design algorithm, which treats them similar to protein side-chain rotamers, except this is for RNA bases. The values are specific to ribose pucker and base identity. This design software can now be used to account for RNA base flexibility while designing RNA molecules, RNA–Protein interfaces, and RNA-small molecule interfaces. Applications of this software include computational mutagenesis studies and structure correction of RNA-Protein interfaces.
DEVELOPMENT OF A PARTICLE SWARM OPTIMIZATION VARIATION OF AUTODOCK FOR EFFICIENT AND ACCURATE DOCKING OF HIGHLY FLEXIBLE LIGANDS

Jorge LS Pina1, Beatriz SLP Lima1, Solange Guimarães2 and Floriano P Silva Junior3

1Laboratório de Bioquímica de Proteínas e Peptídeos, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, Brasil; 2Departamento de Engenharia Civil, UFRJ, Rio de Janeiro, Brasil. *floriano@ioe.fiocruz.br

In this work, six different variants of the standard PSO algorithm were embedded in the AutoDock3 package and tested with distinct topologies and sets of parameters for molecular docking of 10 complexes, ranging from zero to twenty rotatable bonds in the ligand.

INTRODUCTION

Many biologically active ligands, such as peptides and peptidomimetics, have high conformational diversity, i.e., flexibility.1 For instance, an octapeptide may easily have 30 or more rotatable bonds, depending on its amino acid sequence. Currently, the practical means to handle such critical docking cases is to fix the backbone torsions and let only side chain dihedrals free to be optimized.2 Hence, it is of utmost importance to develop docking methods capable of efficiently dealing with highly flexible ligands, i.e., >25 rotatable bonds. The work presented here builds upon the work of Namavivayam & Gunther2 as other variations of the Particle Swarm Optimization (PSO) algorithm are embedded in the AD3 and explored for developing an efficient and accurate tool for docking of highly flexible ligands.

METHODS

The following variants of the standard PSO algorithm were then embedded in AD3: (i) CWPSO – Constant inertia weight factor PSO; (ii) VWPSO – Variable inertia weight factor PSO; (iii) CPSO – Constriction PSO; (iv) RBPSO – Random pbest PSO; (v) FIPSO – Fully-Informed PSO; and (vi) DEPSO – Hybrid differential evolution and PSO. For all six variants considered in this work the number of particles in the swarm (S) was considered as S = 60 + (int) (2 * sqrt (D)), where D is the dimension of the swarm particle, and D = 7 + ntor, where ntor is the number of rotational bonds of the ligand. Four different topologies were tested: (i) Global - All the particles of the swarm contribute to the gbest, which is used by all particles; (ii) Ring - Only the particles with indexes below and above contribute to the gbest to be used by each particle; (iii) Random (3 informants) - Every particle communicates randomly with three other to obtain the gbest to be considered for it. The particle itself is considered part of the group; (iv) Random (variable informants) - Every particle communicates randomly with a percentage of particles in the swarm. The average pbest of the group is the gbest to be used by each particle belonging to this group. The particle itself is not considered in the group. This topology was used only with the FIPSO variant. Six sets of parameters w, c1 and c2 were investigated in this work (Table 1).

PARTICULAR combinations of a variant of the PSO algorithm with a type of topology and a set of parameters were called implementations, making up a total of 46 tested in this work.

RESULTS & CONCLUSIONS

Depending on the parameters used in each of the PSO variants all implementations were capable of giving final ligand configurations comparable to those obtained with the LGA algorithm. Remarkably, above 13 rotatable bonds all variants showed better results than LGA, that is, configurations with lower energies. The variant FIPSO (Fully-Informed PSO) presented RMSD and energy values comparable to or better than CPSO (Constriction PSO) and DEPSO (a hybrid combining Differential Evolution and PSO) while having a better convergence than all other implementations, making it a valuable option for large scale molecular docking, especially for highly flexible ligands such as peptides (Figure 1).

REFERENCE


TABLE 1. Parameter sets for PSO.

<table>
<thead>
<tr>
<th>Set</th>
<th>w1</th>
<th>w2</th>
<th>c1</th>
<th>c2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>6.05</td>
<td>6.05</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>2.8</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>1.02</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
<td>0.65</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>5.05</td>
<td>5.05</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
<td>4.3</td>
<td>4.3</td>
<td></td>
</tr>
</tbody>
</table>

* Set #1 and #2 were obtained from references 1,2. Sets #3-6 were obtained in this work. In the case of the VWPSO variant it was used the range 0.9 to 0.4 for w.
We have subjected several sets of real and simple, computationally generated model proteins to Delaunay tessellation and have computed statistics on both Delaunay simplex geometry and the tendency of quadruplets of residue types to be joined together in simplices. We have characterized the geometry and contact patterns of real proteins and some of the ways in which they differ from these model structures. We have found heretofore unreported asymmetries in contact patterns among residue quadruplets joined in simplices in real proteins.

**INTRODUCTION**

Protein structures have been analyzed with a technique from computational geometry known as Delaunay tessellation\(^1\). In summary, each amino acid is abstracted to a point. These points are then joined by edges, in a unique way, to form a set of non-overlapping, irregular, space-filling tetrahedra (also called simplices). With Delaunay tessellation, we have analyzed five sets of structures, one set of real proteins and four control sets of computationally generated models, compiling statistics on contact patterns in sequence and between amino acid types.

**METHODS**

We have analyzed statistics from the tessellation of five sets of structures: one set of real proteins and four control sets of models. The first set, 1364culled, consisted of the x-ray structures of 1364 non-homologous (30% level) protein chains obtained from the PISCES web server\(^2\). The second dataset, *random-strands* (random polymers each confined to a sphere of diameter 7.177(N\(^{1/3}\) + 2 in order to approximately match the size and shape of globular proteins. There is a counterpart to each 1364culled structure with the same amino acid sequence in random-strands. A third set, *artificial-helices*, was 101 computationally generated, perfectly straight alpha-helices ranging from 50 to 500 residues. A fourth data set, *random-spheres*, was composed of sets of points uniformly distributed in a sphere of diameter 7.177(N\(^{1/3}\) + 2. Residue numbers were assigned randomly so residues close in ‘primary sequence’ were not necessarily close in space. As with *random-strands*, there was one corresponding structure for each real protein in 1364culled. The final set, 1364culled_permuted, were structures from 1364culled with the native sequences randomly permuted.

Consider a Delaunay simplex with residues numbered r1, r2, r3, and r4 at its vertices with r1 < r2 < r3 < r4. A triple of positive integers d1=r2-r1, d2=r3-r2 , and d3=r4-r3 can be associated with each simplex which specifies the spacing in primary sequence of the vertex residues and such triples have formed the basis of a structure comparison algorithm\(^3\). We will refer to a particular triple (d1,d2,d3) as a *d-triple*.

Denote the quadruplet of amino acids at the vertices of a simplex listed in primary sequence order as an *s-ordered quadruplet* and listed in alphabetical order as an *a-ordered quadruplet*. In general, a single *a-ordered quadruplet* will correspond to a family of *s-ordered quadruplets*, namely all distinct permutations of the *a-ordered quadruplet*. For example, the *a*-ordered quadruplet KlLV corresponds to the *s*-ordered quadruplets KlLV, KlVL, KvLL, LkLV, LkVL, LkVL, LkVL, VlKL, VlKL, and VlLK. With a twenty-letter amino acid alphabet, there are 160,000 possible *s*-ordered quadruplets and 8,855 possible *a*-ordered quadruplets.

We have computed the frequencies of *d*-triples and *a*-ordered and *s*-ordered quadruplets for our five sets of structures, and have found patterns present in real proteins that are not present in the models.

**RESULTS & CONCLUSIONS**

Certain *d*-triples occur far more frequently than others in real proteins and artificial helices and to a lesser extent in random strands. This is due in part to some *d*-triples being associated with secondary structure\(^4\). For example (1,1,1), (1,2,1), and (3,1,3) are strongly associated with alpha helices and there are far more of these in real proteins than in random self-avoiding polymers. With random-strands, there is no preference between (n,1,1) and (1,1,n) – both have similar counts. With 1364culled (real proteins) however, there is a strong preference for (n,1,1) over (1,1,n).

One might guess that all *s*-ordered quadruplets occur with equal frequency, however for a substantial minority of *a*-ordered quadruplets, 1D chi-square tests indicate that the corresponding *s*-ordered quadruplets occur with significantly different frequencies\(^5\). For example in the case of *a*-ordered KlLV, the hydrophobic residues tend to prefer positions 1 and 4 in the Delaunay tetrahedron, while lysine prefers positions 2 and 3. This is a remarkable result, since there is no fixed relationship in sequence between the four residues constituting an *s*-ordered quadruplet.

**REFERENCES**

INTRODUCTION

KB-Role is a web resource for the retrieval of predictions of protein functions for protein structures that is available at the URL http://protein.tcmedc.org/KB-Role. The characteristics that are used to make the predictions are diverse and include attributes assigned to the protein’s sequence and/or three-dimensional structure. These characteristics were assembled from the Protein Structure Initiative’s Structural Biology Knowledgebase (PSI SBKB), the UniProt resource, and open biomedical databases. Examples attributes include protein domain assignments, the presence of structural cluster assignment, assignments to cellular and biochemical pathways, disease associations, small molecule interactions, post-translational modifications, and participation in protein-protein interactions. In collaboration with the SIFTS project team, those attributes associated with specific residues within proteins’ sequences were mapped to residues in the protein structures in a comprehensive manner. That was done ensure that they fall within the sequence range for which a three-dimensional structure is available. Collectively the attributes were used to predict Gene Ontology terms and associations with small molecule found in DrugBank. The predictions include possible extensions to the functional assignments of previously characterized proteins and novel predictions of the roles of proteins of unknown function. The aim of the resource is to provide users with hypotheses as to the functional roles of proteins structures of their interest to further guide experiments. As a gauge of the strength of each prediction, the probability of each predicted assignment was estimated with values that ranged from 0 to 1. This work was supported in part by a general funds provided by The Commonwealth Medical College and by a sub-contract of the PSI Structural Biology Knowledgebase from the National Institute of General Medical Sciences (U01 GM093324).

METHODS

Attributes were assembled for structures in the Protein Data Bank at the level of the structural chain as previously described [1]. Additional attributes that were used for the current iteration of the KB-Role prediction set include sequence annotation features as provided in UniProt [2]. Annotations relevant to the region of the protein where the three-dimensional structure is available was found using the mapping files that go from residue level assignments associated with the protein sequence to the residues associated with the protein structure. The mapping between residue numbers in UniProt to residue numbers in PDB were made available through the SIFTS project [3]. Classification models that discern between protein structures with and without a given GO assignment were made as previously described [1]. An extension of the method was to create classification models that discern between small molecules that are described in the DrugBank resource [4].

RESULTS & CONCLUSIONS

The KB-Role website hosts the predictions of Gene Ontology assignments and small molecule associations (Figure 1). The user can search by PDBID and GO term ID for predictions of GO term assignments. To retrieve predictions of small molecule associations, a search may be conducted with PDBID, InChIKey, or a generic drug name. For each prediction, the relevant attributes of the protein structure that were used to make the prediction are presented. These provide the user with a better understanding of what were the important parameters used to make the prediction. Predictions above an estimated probability value of 0.95 are displayed interactively on the pages. All predictions with their associated probabilities may be downloaded from the KB-Role pages described on the left menu of the site.

REFERENCES

Knowledge-based Potential for Positioning Membrane Associated Structures and Assessing Residue Specific Energetic Contributions for Prediction and Design

Chaim A. Schramm1, Brett T. Hannigan2, Jason E. Donald3,6, Chen Keasar5, Jeffrey G. Sevan1, William F. DeGrado1,4,5, Ilan Samish7

1Graduate Group in Biochemistry and Molecular Biology, 2Graduate Group in Genomics and Computational Biology, 3Dept of Biochemistry and Biophysics, School of Medicine and 4Dept of Chemistry, School of Arts and Sciences, 5U. of Pennsylvania, Philadelphia, PA, USA, 6Dept of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva, Israel, 7Current addresses: 4Arividia, Inc., Medford, MA. 5Dept of Pharmaceutical Chemistry, School of Pharmacy, U. of California, San Francisco, San Francisco, CA, USA, 6Dept of Structural Biology, Weizmann Institute of Science, Rehovot, Israel

We produce a non-redundant database to compute knowledge-based asymmetric cross-membrane potentials from the distributions of Cα, Cβ, and functional group atoms1. We predict transmembrane (TM) and peripheral regions from sequence and position structures relative to the bilayer (http://www.degradolab.org/ez). The topological landscapes underscore positional stability and functional mechanisms demonstrated for antimicrobial peptides, TM proteins, and viral fusion proteins. Moreover, experimental point mutations affecting dual-topology proteins are quantitatively reproduced. The functional group potential and the membrane-exposed residues display the largest energetic changes enabling to detect native-like structures from decoys. Hence, focusing on the uniqueness of membrane-associated proteins and peptides, we parameterize their cross-membrane propensity facilitating refinement, characterization, prediction and design.

Introduction

Membrane proteins account for over a quarter of the proteome and most drug targets though constitute only 2% of PDB structures. Computational analysis, prediction and design provide a way to bridge this gap. Membrane-protein interactions are pivotal in folding, stability and energetics. While several rules have been elucidated a non-biased1 analysis and utilization of this relationship is long missing.

Methods

The dataset includes helical structures with resolution < 3.5Å and a sequence identity of >30%. Preference was given to 3D quality (resolution + Rfree), wild-type, mesophile, ground-state and non-engineered structures. Databases were computed for buried and exposed residues by rolling a 1.9 Å-radius lipid methylene probe on the protein. The propensity and effective energy were calculated using reverse Boltzmann statistics data fitting sigmoidal, gaussians and the combination of both where needed. In sequence mode, a 26-residue sliding window was threaded on a helix which was searched for minimum energy location with TM segments forced to alternate topology. In surface mode, a 7-residue window was searched on both membrane sides. Full-genome biases where assigned for confidence score>0.95. In 3D mode a search was carried for HELANAL recognized helices deriving pseudo-energy topological landscapes. For in/out EmrE dual-topology back-calculated Boltzmann probability was computed for the two topologies. Decoys were generated by random rotation around the χ1 dihedral angle followed by application of MESH1 and/or ROSETTA-membrane.

Results

The large representative dataset and use of efficient preprocessing such as lookup tables enables to accurately position peripheral and TM structures and predict the topology of membrane proteins from full genomic sequences1. In sequence mode the results are comparable to complex machine-learning approaches. In 3D mode, the high-resolution of the results enables to quickly produce topological pseudo-energy landscapes distinguishing e.g. between antimicrobial and lytic peptides presenting mechanistic implications and allowing for quick mutational or protein-design scans. Experimental point mutations affecting the topological distribution of dual topology proteins are precisely recapitulated. Moreover, while ROSETTA-membrane can optimize decoy structures focusing on functional-group atoms, our method is more sensitive in selecting the most-native like structure from a decoy set. Hence, along with computational analysis of kinks1, B-factor flexibility1, and motion1 membrane-associated proteins are better understood enabling their analysis, prediction and design.

References

# List of participants

<table>
<thead>
<tr>
<th>First name</th>
<th>Last name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abul</td>
<td>Ullah</td>
<td>SRRC, ARS, USDA</td>
</tr>
<tr>
<td>Ada</td>
<td>Yonath</td>
<td></td>
</tr>
<tr>
<td>Adam</td>
<td>Godzik</td>
<td>Sanford-Burnham Medical Rese</td>
</tr>
<tr>
<td>Alberto</td>
<td>Roca</td>
<td>ProfileGrid.org</td>
</tr>
<tr>
<td>Alexandr</td>
<td>Goncarenco</td>
<td>University of Bergen</td>
</tr>
<tr>
<td>Alpan</td>
<td>Raval</td>
<td>D. E. Shaw Research</td>
</tr>
<tr>
<td>ANA VICTORIA</td>
<td>CONESA CEGARRA</td>
<td>PRINCIPE FELIPE RESEARCH CENT</td>
</tr>
<tr>
<td>Andreas</td>
<td>Prlic</td>
<td>UCSD</td>
</tr>
<tr>
<td>Andreas</td>
<td>Schueller</td>
<td>Pontificia Universidad Catolica de Chile</td>
</tr>
<tr>
<td>Andrew</td>
<td>Borden</td>
<td>Mayo Clinic</td>
</tr>
<tr>
<td>Angelo</td>
<td>Facchiano</td>
<td>Italian National Research Council (CNR)</td>
</tr>
<tr>
<td>ANNA</td>
<td>VANGONE</td>
<td>University of Salerno</td>
</tr>
<tr>
<td>Anna</td>
<td>Marabotti</td>
<td>National Research Council</td>
</tr>
<tr>
<td>Arthur</td>
<td>Olson</td>
<td>The Scripps Research Institut</td>
</tr>
<tr>
<td>Arun</td>
<td>Konagurthu</td>
<td>Monash University</td>
</tr>
<tr>
<td>Avner</td>
<td>Schlessinger</td>
<td>UCSF</td>
</tr>
<tr>
<td>Barry</td>
<td>Grant</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Bastiaan</td>
<td>van den Berg</td>
<td>Delft University of Technolog</td>
</tr>
<tr>
<td>Bradley</td>
<td>Hintze</td>
<td>Duke University</td>
</tr>
<tr>
<td>Brian</td>
<td>Olson</td>
<td>George Mason University</td>
</tr>
<tr>
<td>Bruggmann</td>
<td>Remy</td>
<td>University of Bern</td>
</tr>
<tr>
<td>Byungkook (BK)</td>
<td>Lee</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>Charlotte</td>
<td>Deane</td>
<td>Oxford University</td>
</tr>
<tr>
<td>Chin-Hsien (Emily)</td>
<td>Tai</td>
<td>NCI/NIH</td>
</tr>
<tr>
<td>Chris</td>
<td>Sander</td>
<td>Memorial Sloan-Kettering Cancer Center</td>
</tr>
<tr>
<td>Christine</td>
<td>Orengo</td>
<td>University College London</td>
</tr>
<tr>
<td>Christophe</td>
<td>Blanchet</td>
<td>IDB CNRS</td>
</tr>
<tr>
<td>Constance</td>
<td>Jeffery</td>
<td>University of Illinois at Chi</td>
</tr>
<tr>
<td>Daisuke</td>
<td>Kihara</td>
<td>Purdue University</td>
</tr>
<tr>
<td>Daniel</td>
<td>Keedy</td>
<td>Duke University</td>
</tr>
<tr>
<td>Daryi</td>
<td>Wang</td>
<td>Academia Sinica</td>
</tr>
<tr>
<td>Daven</td>
<td>Vasishtan</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>David</td>
<td>Goodsell</td>
<td>The Scripps Research Inst</td>
</tr>
<tr>
<td>Dietlind</td>
<td>Gerloff</td>
<td>University of California, Santa Cruz 110</td>
</tr>
<tr>
<td>First name</td>
<td>Last name</td>
<td>Affiliation</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>Dominic</td>
<td>Duchene</td>
<td>Universite de Sherbrooke</td>
</tr>
<tr>
<td>Dorota</td>
<td>Latek</td>
<td>International Institute of Molecular &amp; Cell Biology</td>
</tr>
<tr>
<td>Elizabeth</td>
<td>Eskow</td>
<td>University of Colorado, Boulder</td>
</tr>
<tr>
<td>Feng</td>
<td>Zhao</td>
<td>Toyota Tech Inst at Chicago</td>
</tr>
<tr>
<td>Firas</td>
<td>Khatib</td>
<td>UW</td>
</tr>
<tr>
<td>Floriano</td>
<td>Silva Junior</td>
<td>FIOCRUZ</td>
</tr>
<tr>
<td>Francis</td>
<td>Gaudreault</td>
<td>University of Sherbrooke</td>
</tr>
<tr>
<td>Francois</td>
<td>Major</td>
<td>University of Montreal</td>
</tr>
<tr>
<td>George</td>
<td>Nicola</td>
<td>UCSD</td>
</tr>
<tr>
<td>Goran</td>
<td>Neshich</td>
<td>EMBRAPA - CNPTIA</td>
</tr>
<tr>
<td>Guido</td>
<td>Capitani</td>
<td>Paul Scherrer Institut</td>
</tr>
<tr>
<td>Gustavo</td>
<td>Parisi</td>
<td>Universidad Nacional de Quilmes</td>
</tr>
<tr>
<td>Hannah</td>
<td>Edwards</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Hariantto</td>
<td>Tjong</td>
<td>University of Southern California</td>
</tr>
<tr>
<td>Hedi</td>
<td>HegyI</td>
<td>Inst of Enzymology</td>
</tr>
<tr>
<td>Henry</td>
<td>Wilman</td>
<td>Oxford University</td>
</tr>
<tr>
<td>Henry</td>
<td>van den Bedem</td>
<td>Stanford Synchrotron Radiation Lightsource / Structural Genomics</td>
</tr>
<tr>
<td>Ilan</td>
<td>Samish</td>
<td></td>
</tr>
<tr>
<td>Inbal</td>
<td>Budowski-Tal</td>
<td>Technion, Haifa U</td>
</tr>
<tr>
<td>Inbal</td>
<td>Sela-Culang</td>
<td>Bar-Ilan University</td>
</tr>
<tr>
<td>Irene</td>
<td>Farabella</td>
<td>Institute of Structural and Molecular Biology</td>
</tr>
<tr>
<td>Izabella</td>
<td>Neshich</td>
<td>UNICAMP - Universidade Estadual de Campinas</td>
</tr>
<tr>
<td>Jamie</td>
<td>Hill</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Jesse</td>
<td>Eickholt</td>
<td>University of Missouri-Columbia</td>
</tr>
<tr>
<td>Jian</td>
<td>Peng</td>
<td>Toyota Technological Institute at Chicago</td>
</tr>
<tr>
<td>Jianzhu</td>
<td>Ma</td>
<td>Toyota Technological Institute at Chicago</td>
</tr>
<tr>
<td>John</td>
<td>Kim</td>
<td>Rutgers University</td>
</tr>
<tr>
<td>John</td>
<td>Westbrook</td>
<td>Rutgers University</td>
</tr>
<tr>
<td>John-Marc</td>
<td>Chandonia</td>
<td>Berkeley National Lab</td>
</tr>
<tr>
<td>Jose Augusto</td>
<td>Salim</td>
<td>University of Campinas</td>
</tr>
<tr>
<td>Juan</td>
<td>Esquivel</td>
<td>Purdue University</td>
</tr>
<tr>
<td>Kam</td>
<td>Zhang</td>
<td>RIKEN (The Institute of Physical and Chemical Research)</td>
</tr>
<tr>
<td>Keren</td>
<td>Lasker</td>
<td>Stanford</td>
</tr>
<tr>
<td>Kiran</td>
<td>Mukhyala</td>
<td>Genentech, Inc.</td>
</tr>
<tr>
<td>Lazaros</td>
<td>Mavridis</td>
<td>University of St Andrews</td>
</tr>
<tr>
<td>First name</td>
<td>Last name</td>
<td>Affiliation</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>Lenore</td>
<td>Cowen</td>
<td>Tufts University</td>
</tr>
<tr>
<td>Lin</td>
<td>Yang</td>
<td>University of Southern California</td>
</tr>
<tr>
<td>Lloyd</td>
<td>Allison</td>
<td>Monash University</td>
</tr>
<tr>
<td>Marcin</td>
<td>Skwark</td>
<td>Stockholm University</td>
</tr>
<tr>
<td>Mark</td>
<td>Brandt</td>
<td>Rose-Hulman Institute of Technology</td>
</tr>
<tr>
<td>Mark</td>
<td>Wass</td>
<td>CNIO</td>
</tr>
<tr>
<td>Markus</td>
<td>Wiederstein</td>
<td>University of Salzburg</td>
</tr>
<tr>
<td>Martin</td>
<td>Mechelke</td>
<td>MPI for Developmental Biology</td>
</tr>
<tr>
<td>Masahito</td>
<td>Ohue</td>
<td>Tokyo Institute of Technology</td>
</tr>
<tr>
<td>Matsuyuki</td>
<td>Shirota</td>
<td>Tohoku University</td>
</tr>
<tr>
<td>Matthew</td>
<td>Oates</td>
<td>The University of Bristol</td>
</tr>
<tr>
<td>Matthieu</td>
<td>Chartier</td>
<td>Universite de Sherbrooke</td>
</tr>
<tr>
<td>Maya</td>
<td>Topf</td>
<td>Birkbeck College</td>
</tr>
<tr>
<td>Michael</td>
<td>Dolan</td>
<td>NIH</td>
</tr>
<tr>
<td>Mickey</td>
<td>Kosloff</td>
<td>University of Haifa</td>
</tr>
<tr>
<td>Min</td>
<td>Xu</td>
<td>University of Southern California</td>
</tr>
<tr>
<td>Nils</td>
<td>Drechsel</td>
<td>Universitat Pompeu Fabra</td>
</tr>
<tr>
<td>Nir</td>
<td>Ben-Tal</td>
<td>Tel Aviv University</td>
</tr>
<tr>
<td>Noah</td>
<td>Daniels</td>
<td>Tufts University</td>
</tr>
<tr>
<td>Olivier</td>
<td>TABOUREAU</td>
<td>Technical University Denmark</td>
</tr>
<tr>
<td>Patrick</td>
<td>Moss</td>
<td>Alder Biopharmaceuticals</td>
</tr>
<tr>
<td>Patrik</td>
<td>Björkholm</td>
<td>Stockholm University</td>
</tr>
<tr>
<td>Paul</td>
<td>Flook</td>
<td>Accelrys</td>
</tr>
<tr>
<td>Paula</td>
<td>Petrone</td>
<td>Hoffmann-La Roche</td>
</tr>
<tr>
<td>Peter</td>
<td>Rose</td>
<td>University of California, San Diego</td>
</tr>
<tr>
<td>Philip</td>
<td>Bourne</td>
<td>Univ of California, San Diego</td>
</tr>
<tr>
<td>Rafael</td>
<td>Najmanovich</td>
<td>Universite de Sherbrooke</td>
</tr>
<tr>
<td>Red</td>
<td>Crimsonrose</td>
<td>Big Honkin Test Inc</td>
</tr>
<tr>
<td>Remo</td>
<td>Rohs</td>
<td>Univ. of Southern California</td>
</tr>
<tr>
<td>Robert</td>
<td>Fieldhouse</td>
<td>University of Guelph</td>
</tr>
<tr>
<td>Sampada</td>
<td>Bagai</td>
<td>University of Alberta</td>
</tr>
<tr>
<td>Sandhya</td>
<td>Tiwari</td>
<td>University of Bergen, Norway</td>
</tr>
<tr>
<td>Sandor</td>
<td>Vajda</td>
<td>Boston University</td>
</tr>
<tr>
<td>Sebastian</td>
<td>Kelm</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Sheng</td>
<td>Wang</td>
<td>TTIC</td>
</tr>
<tr>
<td>Shruthi</td>
<td>Viswanath</td>
<td>University of Texas at Austin</td>
</tr>
<tr>
<td>Shruti</td>
<td>Koulgi</td>
<td>Centre for Development of Advanced Computing</td>
</tr>
<tr>
<td>First name</td>
<td>Last name</td>
<td>Affiliation</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>Shula</td>
<td>Shazman</td>
<td>Columbia-Presbyterian Medical Center</td>
</tr>
<tr>
<td>Sikander</td>
<td>Hayat</td>
<td>Stockholm University</td>
</tr>
<tr>
<td>Spencer</td>
<td>Bliven</td>
<td>University of California, San Diego</td>
</tr>
<tr>
<td>Stefan</td>
<td>Senn</td>
<td>Rutgers the State University of New Jersey</td>
</tr>
<tr>
<td>Stephen</td>
<td>MacKinnon</td>
<td>Hospital for Sick Children</td>
</tr>
<tr>
<td>Steven</td>
<td>Darnell</td>
<td>DNASTAR, Inc.</td>
</tr>
<tr>
<td>Swati</td>
<td>Jain</td>
<td>Duke University</td>
</tr>
<tr>
<td>Thomas</td>
<td>Haschka</td>
<td>Université de Reims Champagne Ardenne</td>
</tr>
<tr>
<td>Tianyin</td>
<td>Zhou</td>
<td>USC</td>
</tr>
<tr>
<td>Tianyun</td>
<td>Liu</td>
<td>Stanford University</td>
</tr>
<tr>
<td>Timothy</td>
<td>Nugent</td>
<td>University College London</td>
</tr>
<tr>
<td>Tom</td>
<td>Ferrin</td>
<td>Univ. of California, San Fran</td>
</tr>
<tr>
<td>Tomas</td>
<td>Norambuena</td>
<td>Pontificia Universidad Catolica de Chile</td>
</tr>
<tr>
<td>Torsten</td>
<td>Schwede</td>
<td>SIB Swiss Institute of Bioinformatics</td>
</tr>
<tr>
<td>Vered</td>
<td>Kunik</td>
<td>Bar-Ilan University</td>
</tr>
<tr>
<td>Vincent</td>
<td>Frappier</td>
<td>Universite de Sherbrooke</td>
</tr>
<tr>
<td>Yang</td>
<td>Shen</td>
<td>Toyota Technological Institute at Chicago</td>
</tr>
<tr>
<td>Yoonjoo</td>
<td>Choi</td>
<td>Dartmouth College</td>
</tr>
<tr>
<td>Yosi</td>
<td>Shibberu</td>
<td>Rose-Hulman Institute of Technology</td>
</tr>
<tr>
<td>Yuri</td>
<td>Matsuzaki</td>
<td>Tokyo Institute of Technology</td>
</tr>
</tbody>
</table>