

Computation Resources for Molecular Biology: Special issue 2017

Marina I Ostankovitch (1) and Michael J E Sternberg (2)*

(1) Journal of Molecular Biology
50 Hampshire St
Cambridge MA USA

(2) Structural Bioinformatics Group
Centre for Integrative systems Biology and Bioinformatics
Department of Life Sciences
Imperial College London
South Kensington
London SW7 2AZ UK

(*) To whom communication should be sent - m.sternberg@imperial.ac.uk

We are pleased to present the second edition in JMB of a Special Issue that highlights computation resources for molecular biology. Our aim is to report both new developments and substantial improvements to proven existing resources. We feature both web-based resources, which provide easy access to users, and downloadable tools of particular use for in-house processing or the inclusion into pipelines being developed in the laboratory.

The number of determined protein structures continues to increase and highly-curated databases are required to provide an organized view of this highly complex information. Brenner and colleagues present Structural Classification of Proteins—extended (SCOPe) that is a database of relationships between protein structures that extends the Structural Classification of Proteins (SCOP) database [1]. SCOP is a fully manually and expert curation of domains from known protein structures arranged in a hierarchical order based on structural and evolutionary relationships. SCOPe primarily uses precise automated tools to extend in a highly accurate and rigorous manner the classification of SCOP that was released in 2009 but in addition uses manual curation to resolve difficult assignment problems. In addition, SCOPe incorporates and improves data bases and tools, like ASTRAL compendium, that aid in the analysis of the sequences and structures of proteins classified in SCOPe. SCOPe 2.06 contains 77,439 Protein Data Bank entries, double the 38,221 structures classified in SCOP [1].

Antibodies are a major family of protein with wide ranging impact on medicine including the design of novel biological therapies. Martin and colleagues report abYsis, a web interface for antibody sequence and structure analysis that integrates a number of applications previously developed by the authors. Searches can be conducted using text, sequence, or structural constraints [2]. abYsis annotates the antigen-binding site following established conventions and numbering systems, identifies potential post-translational modifications, and provides residue frequency for each position in an antibody. This interface provides information that can guide scientists who engineer antibodies on deciding which residues to modify.

A major theme in this year's is the modelling and understanding of protein complexes and their interactions. Blundell and colleagues present Arpeggio, a web server that integrates approaches and computes molecular interactions within and between proteins and protein, DNA, or small-molecule ligands [3]. An important strength of Arpeggio is that it readily displays the data via a simple-to-use click-and-go interface. Moreover it readily provides the users with comprehensive information on interactions which typically are not often available with many other programs (e.g. cation- π , atom-aromatic).

Vajda and colleagues present ClusPro-DC that aims at determining the oligomeric state of a protein with biological function based on crystallographic data [4]. ClusPro is a widely-used resource to predict the docking of two proteins. ClusPro-DC is an extension that classifies dimer upon energy landscape by exhaustively samples the energy landscape during docking between two protein subunits. One can predict that dimers have a biological function when low energy docking poses cluster in the vicinity of the native structure of the dimer and make the interaction stable, thus putatively biologically relevant. If the cluster around the native structure is not large enough, the subunits are unlikely to form a stable dimer and show biological function. ClusPro-DC is made accessible via the existing docking server ClusPro.

Oliva and colleagues present iFrag, a web server that predicts possible interacting regions, even if they are only of a few residues, between pairs of proteins, by searching for minimal sequence similar to regions of other pairs of proteins known to interact [5]. iFrag does not require three-dimensional structural information or multiple sequence alignments in contrast to current methods. The authors present experimental validation of their approach on the successful prediction of a peptide competing with the aggregation of beta-amyloid in Alzheimer's disease.

The stereochemical constraints on single-spanning sections of transmembrane proteins provide the opportunity to develop enhanced resources focused on these molecules. Pogozheva and Lomize have developed TMDOCK, a webserver that hosts an effective computational method for modeling the formation of dimer between single-spanning transmembrane proteins [6]. A sequence of interest is threaded over a limited set of templates of known structure to generate a starting configuration which is then minimized using a customized energy function to generate the model of the docked pair. This new method identifies α -helix dimerization modes, explores the range of possible solutions through multiple rounds of calculations, and ranks them based on the calculated free energy of α -helix association.

Increasingly multi-chain macromolecular complexes are being studied. Bonvin and colleagues present recent advances and web servers of two computational tools PowerFit and DisVis that integrate structural experimental data and provide a structural interpretation and molecular model to elucidate the architecture of macromolecular complexes [7]. PowerFit that combines density data sets resulting from Cryo-electron microscopy has been updated to generate reliability indicators of fitting of density maps. DisVis quantifies and identifies false-positive distance restraints obtained from chemical cross-links coupled with mass spectrometry. It now gives prediction on the location of the ligand based on the analysis of the presumed residues at the contact interface.

William Taylor describes applications of the computer program SimGen for modeling complex macromolecules typified by the multi-domain globular protein actin, the transmembrane protein rhodopsin, and the membrane components of the mitochondrial ATPase [8]. This tool is designed to model large molecular systems using a hierarchical decomposition of molecular or cellular systems to

perform simulations (default motion is a Brownian-like diffusion). SimGen has an uncommon large variety of applications from modelling coarse-grained molecules to modelling the collection of cells.

Knowledge about the interactome is providing new insights into biology at the systems level. Chen and colleagues report the update of their PROFEAT server whose capacity has been extended for computing up to 329 network descriptors and protein–protein interaction descriptors [9]. PROFEAT has the capacity to compute many more network descriptors than other tools available (e.g. Cytoscape and igraph). PROFEAT network descriptors describe the topological and connectivity characteristics of unweighted (uniform binding constants and molecular levels), edge-weighted (varying binding constants), node-weighted (varying molecular levels), edge-node-weighted (varying binding constants and molecular levels), and directed (oriented processes) networks. The authors illustrate the usefulness of the network descriptors with studies of genomic, interaction, transcription, metabolic networks and disease profiles reported in the literature.

The rapid decrease in the cost of genome sequencing is providing a vast source of information about the relationship between genetic variation and phenotype. This is central to understanding the molecular basis of disease and accordingly there have been major developments of resources to predict the effect of sequence variants. Zhang and colleagues present BindProfX that significantly improves the previous method BindProf [10]. BindProfX calculates the impact of gene mutations on the binding affinity through the structure-based profiling of protein–protein interfaces, where the binding free-energy change is counted as the logarithm of relative probability of mutant amino acids over wild-type ones in the interface alignment matrix. BindProfX is particularly advantageous for counting the coupling effect of multiple individual mutations and complementing physical measurements. It outperforms other methods and correlates strongly with experimental measurements.

Gerstein and colleagues have developed Intensification that aims at identifying residue positions at the codon level where variants are likely to be associated with disease [11]. The authors use large scale genome sequencing data to analyze intra-species population-genetic signals. The tool generates multiple alignments of protein motifs to identify positions that are conserved within the motifs that may not be identified from using full length sequences and amplify those that have been strongly conserved between species during evolution and in population.

In this Special Issue we also report a statistical software package developed by Xu and colleagues to compare differences in gene expression patterns upon different biological conditions [12]. BFDCA (“Bayes Factor approach for Differential Co-expression Analysis”) is written in the R language and use statistics to integrates various aspects of differential co-expression patterns (including Shift, Cross, and Re-wiring) into one Bayes factor. The authors validated their software using simulation and experimental data. It outperforms existing methods in accuracy and robustness to detect pairs and modules from results obtained from simulations. When using existing clinical data, BFDCA groups effectively genes into functional subunits and predicts the impact on the regulation of genes related to diseases and find novel putative modules related to disease. BFDCA predicts phenotypes by using significant differential gene expression pairs as markers with high accuracy.

We hope that the readers will find this Special Issue helpful in identifying tools to help them in their study of particular biological systems. In addition, this Issue is also providing an insight into current developments in bioinformatics where the articles describe the strategies being employed to provide easy-to-use access to probe biological data at a range of levels of organization and complexity.

1. John-Marc Chandonia, Naomi K. Fox, Steven E. Brenner, SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins – extended Database, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.11.023>.
2. Mark B. Swindells, Craig T. Porter, Matthew Couch, Jacob Hurst, K.R. Abhinandan, Jens H. Nielsen, Gary Macindoe, James Hetherington, Andrew C.R. Martin, abYsis: Integrated Antibody Sequence and Structure—Management, Analysis, and Prediction, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.08.019>.
3. Harry C Jubb, Alicia P Higuero, Bernardo Ochoa-Montaña, Will R Pitt, David B Ascher, Tom L Blundell, Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.12.004>.
4. Christine Yueh, David R. Hall, Bing Xia, Dzmitry Padhorny, Dima Kozakov, Sandor Vajda, ClusPro-DC: Dimer Classification by the Cluspro Server for Protein–Protein Docking, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.10.019>.
5. Javier Garcia-Garcia, Victoria Valls-Comamala, Emre Guney, David Andreu, Francisco J. Muñoz, Narcis Fernandez-Fuentes, Baldo Oliva, iFraG: A protein–protein interface prediction server based on sequence fragments, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.11.034>.
6. Andrei L. Lomize, Irina D. Pogozheva, TMDOCK: An Energy-Based Method for Modeling α -Helical Dimers in Membranes, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.09.005>.
7. G.C.P. van Zundert, M. Trellet, J. Schaarschmidt, Z. Kurkuoglu, M. David, M. Verlato, A. Rosato, A.M.J.J. Bonvin, The DisVis and PowerFit web servers: Explorative and Integrative Modeling of Biomolecular Complexes, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.11.032>.
8. William R. Taylor, SimGen: A General Simulation Method for Large Systems, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.10.011>.
9. P. Zhang, L. Tao, X. Zeng, C. Qin, S.Y. Chen, F. Zhu, S.Y. Yang, Z.R. Li, W.P. Chen, Y.Z. Chen, PROFEAT Update: A Protein Features Web Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.10.013>.

10. Peng Xiong, Chengxin Zhang, Wei Zheng, Yang Zhang, BindProfX: Assessing mutation-induced binding affinity change by protein interface profiles with pseudo counts, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.11.022>.
11. Jieming Chen, Bo Wang, Lynne Regan, Mark Gerstein, Intensification: A resource for amplifying population-genetic signals with protein repeats, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.12.003>.
12. Duolin Wang, Juexin Wang, Yuexu Jiang, Yanchun Liang, Dong Xu, BFDCA: A Comprehensive Tool of Using Bayes Factor for Differential Co-Expression Analysis, *Journal of Molecular Biology*, 2017, <http://dx.doi.org/10.1016/j.jmb.2016.10.030>.