Poster I-59

Prediction of protein complexes using evolutionary information

Authors:
Richard Bickerton (*Department of Biochemistry, University of Cambridge*)
Tom Blundell (*Department of Biochemistry, University of Cambridge*)
David Burke (*Department of Biochemistry, University of Cambridge*)

Short Abstract: We exploit structural and evolutionary information to predict protein-protein interaction sites for use as restraints in data-driven docking. To benchmark docking algorithms we created a comprehensive non-redundant database of observed protein interactions. This is useful in assessing impacts of non-synonymous SNPs on protein interactions.

Long Abstract:

70% of eukaryotic genes work through multiprotein systems. Despite several advances and some excellent recent examples, experimental determination of protein complexes remains hard. For X-ray crystallography crystallization remains a bottleneck and for NMR techniques there is an upper size limit of around 100kDa. As such, fast, reliable, in silico methods for prediction of protein complexes would be highly desirable. Protein-protein docking involves the prediction of the three dimensional structure of a protein complex from the individual structures of its components. The field is now well established with a variety of algorithms published and progress is periodically assessed objectively in the CAPRI experiment. Most of the published algorithms comprise a brute force rigid-body search, expedited using Fast Fourier Transform techniques, with each of the generated poses scored and ranked. The scoring schemes are almost all based on molecular mechanics-type functions. Structural flexibility remains a major challenge; some algorithms can accomodate some local flexibility but large conformational change remains an unsolved problem. In data-driven orguided docking, information from external sources (e.g. mutagenesis studies, NMR chemical shifts, Cryo EM, in silico site prediction) are used as a restraint to reduce the search space or to modify the scoring function. We have developed the application of environment-specific substitution tables (ESSTs) to aid the identification of protein interaction sites. The ESSTs are derived from a set of high quality structural alignments with scores calculated from the rates of observed substitutions in a set of 24 different structural evironments, as defined by secondary structure, solvent accessibility and hydrogen bonding. The program Crescendo is used to identify evolutionary restraints on protein sequence and structure. This is achieved by comparing, for each amino acid position, the sequence conservation observed in the homologous family of proteins with the degree of conservation expected on the basis of amino acid type and local structural environment. This identifies those residues that have a higher degree of conservation than expected and are therefore likely to be involved in interactions. Importantly our method differs from other published techniques by differentiating restraints that arise from structure from those that derive from intermolecular interactions that mediate functions. The predicted interaction sites are then used as a restraint in the guided docking program PyDock, initially developed by Juan Fernandez-Recio. Although the CAPRI experiment provides a valuable way of monitoring progress in the wider field, the number of

complexes included is small and somewhat non-representative. In order to more broadly benchmark our docking algorithms we have created Piccolo: a comprehensive non-redundant database of pairwise protein-protein interactions of known three dimensional structure. Piccolo is built by taking all records in the Protein Data Bank containing more than one chain. For each pair of chains, the Euclidean distance of every atom in the first chain is measured to every atom in the second chain. This is achieved using the CCP4's NCONT program which performs well owing to an internal bricking algorithm; the whole PDB is processed overnight on a typical workstation. All inter-atomic distances of less than 6.05A are recorded; this value was chosen as it is the maximum length of a water-mediated hydrogen bond. Additionally, all contacts where the observed distance is less than the sum of the two atomic radii plus 0.5A are further flagged as close contacts. Spurious non-physiological crystal artifacts are filtered out based on the number of interactions and the size of the interface (calculated as the difference in accessible surface area between that of the two monomers and that of the complex). The same process has been repeated on the EBI's PQS database of quaternary structures, as this ought to contain more physiologically representative protein complexes than those provided in the PDB. An all-by-all BLAST search is performed on the sequences of the PDB chains. This allows hetero-oligomers, homo-oligomers and homologous hetero-oligomers to be automatically distinguished, as well as facilitating the identification of the non-redundant set of interactions. This is possible at a range of definitions of redundancy - from strict identity to detectable sequence similarity. The data is loaded into an SQLite database. In the PDB release of November 2005 there were 15,522 PDB records containing more than one chain, 53,522 pairs of interacting chains 5,653,387 pairs of interacting residues and 72,385,751 pairs of interacting atoms. A web front-end is under development, using the perl Catalyst framework and the Jmol viewer, which will ultimately be made publicly available. Serendipitously the data has been found to have further application to a series of collaborations the group is involved in around non-synonymous Single Nucleotide Polymorphisms (nsSNPs). The group has developed a suite of software for assessing the likely impact of a nsSNP on a protein's structure, function and interactions. As part of this process Crescendo has been used to identify putative protein interaction sites. If a nsSNP can be mapped to a residue that falls within this site, predictions can then be made on the likely impact of the mutation on the interaction site. The data in Piccolo supplements these observations with a smaller but more reliable set of observed interactions. Of particular interest are instances where two or more nsSNPs have been mapped to the same pairwise protein-protein interaction, on the same or on different proteins. This could potentially identify molecular mechanisms of polygenic disease, where mutations act synergistically; each individually does not compromise the interaction and is not phenotypically visible but in combination could have a phenotypic impact. Work is underway to validate these predictions. Such inferences are only possible through investigation at the structural level.