Limsoon's visit to Yang Ming National University, Taipei, June 2006

Talk1

Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions

Most approaches in predicting protein function from proteinprotein interaction data utilize the observation that a protein often share functions with proteins that interacts with it (its level-1 neighbours). However, proteins that interact with the same proteins (i.e. level-2 neighbours) may also have a greater likelihood of sharing similar physical or biochemical characteristics. We speculate that two separate forms of functional association accounts for such a phenomenon, and a protein is likely to share functions with its level-1 and/or level-2 neighbours. We are interested to find out how significant is functional association between level-2 neighbours and how they can be exploited for protein function prediction. We made a statistical study on recent interaction data and observed that functional association between level-2 neighbours is clearly observable. A substantial number of proteins are observed to share functions with level-2 neighbours but not with level-1 neighbours. We develop an algorithm that predicts the functions of a protein in two steps: (1) assign a weight to each of its level-1 and level-2 neighbours by estimating its functional similarity with the protein using the local topology of the interaction network as well as the reliability of experimental sources; (2) scoring each function based on its weighted frequency in these neighbours. Using leave-one-out cross validation, we compare the performance of our method against that of several other existing approaches and show that our method performs well.

Talk 2

Building Gene Networks by Information Extraction, Cleansing, & Integration

Many common diseases in humans are not caused by a genetic variation within a single gene, but are influenced by complex interactions among multiple genes, environmental, and life style factors. The construction of gene networks is therefore helpful in understanding mechanisms of diseases and in designing effective therapies. I will present several example efforts at the Singapore Institute for Infocomm Research in constructing databases of such gene networks. I will also present some of the challenges and advances in information extraction, data cleansing, and data integration that are pertinent to constructing such databases.

Talk 3

Discovering Motif Pairs at Interaction Sites from Protein Sequences on a Proteome-Wide Scale

Protein–protein interaction, mediated by protein interaction sites, is intrinsic to many functional processes in the cell. We propose here a novel method to discover patterns in protein interaction sites. We observed from protein interaction networks that there exist a kind of significant substructures called interacting protein group pairs, which exhibit an all-versus-all interaction between the two protein-sets in such a pair. The full-interaction between the pair indicates a common interaction mechanism shared by the proteins in the pair, which can be referred as an interaction type. Motif pairs at the interaction sites of the protein group pairs can be used to represent such interaction type, with each motif derived from the sequences of a protein group by standard motif discovery algorithms. The systematic discovery of all pairs of interacting protein groups from large protein interaction networks is a computationally challenging problem. We solve the problem using efficient algorithms for mining frequent patterns through an interesting transformation. We found 5349 pairs of interacting protein groups from a yeast interaction dataset. The expected value of sequence identity within the groups is only 7.48%, indicating non-homology within these protein groups. We derived 5343 motif pairs from these group pairs, represented in the form of blocks. Comparing our motifs with domains in the BLOCKS and PRINTS databases, we found that our blocks could be mapped to an average of 3.08 correlated blocks in these two databases. The mapped blocks occur 4221 out of total 6794 domains (protein groups) in these two databases. Comparing our motif pairs with iPfam consisting of 3045 interacting domain pairs derived from PDB, we found 47 matches occurring in 105 distinct PDB complexes. Comparing with another putative domain interaction database InterDom, we found 203 matches.

A Master Class On:

Knowledge Discovery Techniques for Bioinformatics

This short course will present a broad review of knowledge discovery techniques and also several in-depth studies of their applications in bioinformatics. It is intended for final-year students and post-graduate students starting work with general data mining, machine learning, or bioinformatics. No special statistics, mathematics, or computer science background is required. The course is about 6 hours, comprising the following parts:

(1) Essence of Knowledge discovery [1.5 hours]
- Feature generation, feature selection, feature integration
- Understanding accuracy
- Cross validation

(2) Machine learning methods [1.5 hours]
- Decision trees
- Ensemble classifiers
- K-NN
- Naïve Bayes
- HMM

(3) Gene feature recognition [1.5 hours]
- Translation initiation site
- Transcription start site

(4) Gene expression analysis [1.5 hours]
- Gene expression profile classification
- Gene expression profile clustering
- Extreme sample selection

(5) Protein subcellular localization prediction [1.0 hours]
- Sorting signals
- Amino acid & domain compositions

(6) Database cleansing [0.5 hours]
- Association rules mining
- Errors in databases
- Data de-duplication

(7) Sequence homology interpretation [1.5 hours]
- Guilt by association
- Active site & domain discovery
- What if no sequence homology is found
- Key mutation site discovery

Short CV

Limsoon Wong is a professor in the School of Computing and the School of Medicine at the National University of Singapore. Before that, he was the Deputy Executive Director for Research at A*STAR's Institute for Infocomm Research. He is currently working mostly on knowledge discovery technologies and is especially interested in their application to biomedicine. Prior to that, he has done significant research in database query language theory and finite model theory, as well as significant development work in broad-scale data integration systems. Limsoon has written about 100 research papers, a few of which are among the best cited of their respective fields. In recognition for his contributions to these fields, he has received several awards, the most recent being the 2003 FEER Asian Innovation Gold Award for his work on treatment optimization of childhood leukemias. He serves on the editorial boards of *Journal of Bioinformatics and Computational Biology* (ICP), *Bioinformatics* (OUP), and *Drug Discovery Today* (Elsevier). He is a scientific advisor to GeneticXchange (USA), Molecular Connections (India), and KooPrime (Singapore). He received his BSc(Eng) in 1988 from Imperial College London and his PhD in 1994 from University of Pennsylvania.