Predicting Cancer Interaction Networks Using Text-Mining and Structure Understanding

Christopher M. Topinka, Chi-Ren Shyu, Ph.D.
Department of Computer Science, University of Missouri - Columbia, MO 65211

Abstract. Extended biomolecule binding or interaction networks can be built by computationally predicting protein-protein interactions from diverse data sources. To construct networks focused on cancer research our approach combines domain specific natural language processing (NLP) assisted text-mining of biomedical literature databases with structure-based protein-protein interaction prediction reinforced with sub-cellular localization and evolutionary information. Fast retrieval of structure-based queries will be accomplished by using a novel knowledge discovery process developed previously [3].

Background. Understanding "complex diseases" such as cancer requires knowledge of multiple relationships between the function of possibly many interacting cellular components. A system that determines the specifics of protein-protein interfaces as well as accurately predicts a network of putative protein interactions for a given set of target proteins associated with abnormal function for a given disease domain could be of particular use in visualizing potential target sequence sites in drug design and for discovering new levels of complexity of interaction relevant to a specific disease or therapy.

Biological interaction networks can be visualized as a graph where nodes are physical or conceptual biological entities and edges are relationships between entities. NLP assisted text-mining has been used to encode biological networks with edge relevant information such as the catalytic nature of the interaction. Other encodings include the edge weight as the number of documents found in the PubMed biomedical literature database with evidence supporting the edge relationship which can also be interpreted as the relative strength of an interaction through simple co-occurrence of gene names within a document [1, 2]. Where NLP text-mining has proven useful for the production of content-rich relationship networks between biological entities, experimental results from approaches using combined methods have shown that the predictive power of structure-based interaction prediction methods are a more powerful predictor than other information sources [4]. Structural homology prediction can be achieved by threading pair-wise partial alignments of known template interface areas to target protein surfaces. Combined methods benefit from being able to extend and annotate the resulting network in multiple and meaningful ways.

Challenges. Several challenges exist in developing an accurate interaction network prediction tool that addresses the known complexity of relationships and gaps in knowledge present in the cancer research field. Refining comprehensive domain focused term dictionaries extracted from the UMLS, GO, OMIM, HUGO and other existing biomedical term sets is the first task in developing NLP algorithms that can provide network annotation meaningful to the cancer domain. Machine learning techniques can assist in integrating structure-based interaction prediction rankings with non-structure-based evidence into a combined prediction. Flexible and intuitive visualization of the annotated network is critical for incorporating interaction information with other information sources, such as microarray analysis, into a streamlined process available to cancer researchers.

Goal. Provide a Web-based application that accepts structured queries that provide the initial seed for extending a biological network to likely interaction partners. Query result sets derived from the prediction system will include all potential interacting entities, the predicted strength and catalytic influence of putative bindings along with epigenetic control factors affecting the network.

Acknowledgement
CMT is funded by the National Library of Medicine – Biology and Health Informatics Research Training Grant No. 5 T15 LM07089 13. CRS is supported by Shumaker Endowment in Bioinformatics

References

1. Chen, H. & Sharp, B. M. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*. 2004; 5:147.
2. Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*. 2001; 28(1):21-8.
3. Shyu, C.R., Chi, P.H., Scott, G. & Xu, D. ProteinDBS: a real-time retrieval system for protein structure comparison. *Nucleic Acids Res*. 2004; 32 (Web Server issue):W572-5.
4. Singh, R. Struct2Net: Integrating Structure into Protein-Protein Interaction Prediction, *Pacific Symposium on Biocomputing*. 2006; 11:403-414.