MAR05-2004-001965

Protein interaction networks from literature mining

SIGEO IHARA, The University of Tokyo

The ability to accurately predict and understand physiological changes in the biological network system in response to disease or drug therapeutics is of crucial importance in life science. The extensive amount of gene expression data generated from even a single microarray experiment often proves difficult to fully interpret and comprehend the biological significance. An increasing knowledge of protein interactions stored in the PubMed database, as well as the advancement of natural language processing, however, makes it possible to construct protein interaction networks from the gene expression information that are essential for understanding the biological meaning. From the *in house* literature mining system we have developed, the protein interaction network for humans was constructed. By analysis based on the graph-theoretical characterization of the total interaction network in literature, we found that the network is scale-free and semantic long-ranged interactions (i.e. *inhibit*, *induce*) between proteins dominate in the total interaction network, reducing the degree exponent. Interaction networks generated based on scientific text in which the interaction event is ambiguously described result in disconnected networks. In contrast interaction networks based on text in which the interaction events are clearly stated result in strongly connected networks. The results of protein-protein interaction networks obtained in real applications from microarray experiments are discussed: For example, comparisons of the gene expression data indicative of either a good or a poor prognosis for acute lymphoblastic leukemia with *MLL* rearrangements, using our system, showed newly discovered signaling cross-talk.