Identification of gene annotations and interactions and protein-protein interaction associated disorders through data integration

A. Canakoglu, P. Gangi, S. Gennaro, M. Masseroli✉
Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy

Motivations

Available biomolecular annotations are very valuable, but dispersed and far from being complete. High quality integration of scattered annotation data and reliable identification of new annotations can greatly support unveiling new biomedical knowledge. Biomolecular interactions are one of the main objectives of biomolecular studies, due to the understanding that biological processes are mainly driven by interactions among biomolecular entities, such as proteins and DNA. New powerful high-throughput experimental techniques are providing numerous protein-protein interaction data; they are being collected, together with computational results, in several different databases, which include IntAct, BioGrid, BIND, DIP, HPRD and MINT. Yet, they generally do not include phenotypic or even functional or structural information about the interactors, which in many cases are available in other databases. In particular, no information is available about the association of protein-protein interactions with genetic disorders. This creates the need to integrate the sparsely available data in order to enrich the identified interactions with additional evidence, support their biological interpretation and identify their involvement in inherited pathologies. In addition, integration of gene and protein annotations and protein-protein interaction data provides the base to infer new gene annotations and interaction networks.

Methods

We developed an automatic association inference method, based on the transitive closure concept, and applied it on the data from several distributed sources integrated in our Genomic and Proteomic Data Warehouse (GPDW). In particular, by leveraging protein-protein interaction data, provided by the IntAct and MINT databases, and protein encoding gene data form the Entrez Gene database, we inferred gene interaction networks. In addition, by taking advantage of genetic disorder and phenotype data provided by the OMIM database, we inferred associations between proteins and genetic disorders and their phenotypes. Then, in order to identify genetic disorders possibly associated with protein-protein interactions, we looked for those interacting proteins that resulted associated with the same genetic disorder.

Results

Our GPDW currently includes 46,154 human protein-protein interactions (out of the 254,048 protein-protein interactions contained), which involve 12,178 different human proteins (out of the 326,766 human proteins in the GPDW) that are encoded by 11,232 different human genes. By applying the above described method, we identified 1,130 gene networks and found 1,136 human protein-protein interactions associated with 628 genetic disorders (6% of all genetic disorders in the GPDW), which are related to 86 clinical synopses (87% of all clinical synopses in the GPDW) and 3,481 phenotypes (10% of the total phenotypes in the GPDW). Among others, we found four interacting proteins (AHSA1 _ HUMAN, CFTR _ HUMAN, DERL1 _ HUMAN, and RNF5 _ HUMAN) whose encoding genes are known to be associated with cystic fibrosis, as well as other 43 genes. Mutations of the CFTR human gene are known to be directly involved in different grades and manifestations of cystic fibrosis. The found associations of the AHSA1 _ HUMAN, CFTR _ HUMAN, DERL1 _ HUMAN, and RNF5 _ HUMAN interacting proteins with cystic fibrosis could suggest that some types of this multi-variant disorder may be associated with defects in the interactions between these proteins. Possibly, different CFTR _ HUMAN protein mutations could alter its functional interaction with one or more of the AHSA1 _ HUMAN, DERL1 _ HUMAN and RNF5 _ HUMAN proteins. If this would be proven, such finding would also suggest, as a possible disease treatment strategy, the engineering of a synthetic protein interacting with the mutated CFTR _ HUMAN protein and similar in function to the one of the AHSA1 _ HUMAN, DERL1 _ HUMAN or RNF5 _ HUMAN proteins whose interactions with the mutated CFTR _ HUMAN pro-

tein result altered. The above discussed findings demonstrate the importance of the transitive closure based inference method developed, as well as of the data integration approach implemented in the GPDW and the relevance of the comprehensive data there integrated. The GPDW constitutes the backend of a Genomic and Proteomic Knowledge Base (GPKB) that is publicly available at http://www.bioinformatics.dei.polimi.it/GPKB/ through a prototype easy-to-use and efficient Web interface.

Availability
http://www.bioinformatics.dei.polimi.it/GPKB/