Poster B-19

APID: Agile Protein Interaction DataAnalyzer

Authors:
Prieto C. (*Cancer Research Center (CIC, CSIC/USAL)*)
De Las Rivas J. (*Cancer Research Center (CIC, CSIC/USAL)*)

Short Abstract: Nowadays the assessment of the reliability and broader coverage of the interactome network are two of the main research areas in protein interactions. With these purposes APID (http://bioinfow.dep.usal.es/apid/) has been developed to analyze and integrate in a comparative web platform all the currently known data about protein-protein interactions.

Long Abstract:
INTRODUCTION At present time one of the most productive areas of biological data is protein-protein interactions. The data about the interaction of two or more proteins are stored in published scientific papers where the information is difficult to manage and compute. For this reason several bioinformatic initiatives have been undertaken to store, in biological databases, information about protein interactions. These initiatives tend to extract and integrate experimental knowledge about interacting proteins from scientific journals in their database. Each initiative has its own extraction, curation and storage protocols. This is the reason why the intersection and overlap between these source databases is small, and therefore in many cases their information is complementary and can be unified to increase our knowledge about interactomes of different species. But this unification isn't trivial; the lack of a unique protein code, the different protocols to obtain, validate and store the information, the varied annotation about experimental methods and the curator's interpretation about the journal give a lot of heterogeneity between the data inside each database and obstruct the unification of the protein interaction data. Considering all these problems, a web application with a comprehensive unification and integration of the known protein interaction data could be helpful for the research community. DATA UNIFICATION All the work has been developed in Java, and a J2EE architecture has been used to build the web interface and the network browser. The data unification has been done based on three key reference identifiers: (i) UniProt accession code, to allow a specific identification of each protein; (ii) PSI-MI ID, to unify the experimental methods used in different publications to a common terminology developed by PSI-MI; (iii) PubMed ID (PMID), to attach each interaction validated with a given experimental method to a specific PubMed literature reference. These three main key identifiers constitute a simple information core that makes APID an agile web server easy to access and search. At the same time, these three key identifiers allow only to get additional data about the proteins or about the interactions by link to other biological data sources. In this way, we have obtained a protocol able to store and unify protein interaction databases in a clear uniform structure, maintaining the integrity of the data and correcting some existing failures found in the original files. INTERACTION ASSESSMENT In recent years several studies have reported comparative assessments of large-scale and high throughput protein-protein interaction data indicating that data quality is a critical problem in these data sets that many times include a high proportion of false positive interactions due to

low accuracy of the methods. Some bioinformatic and computational work has been done to assess the reliability of high throughput observations and to gain confidence in the data. Following these efforts that consider the strong need to improve the data quality, we have included in our web application certain calculated parameters that weigh the reliability of a given interaction. These parameters are based on the number of experimentally validated methods that prove a protein-protein interaction, the overlap of GO terms in the protein interactors and the interaction of domains that are presented in iPfam as interacting domains. In this way APID is a unified repository of interactions where it is able to validate the interactions inside thorough the parameters described above. As well as give information about interactions, APID also infer data about proteins in the interaction network, the web application calculate graph parameters as the connectivity and the cluster coefficient of each node, and a functional prediction of each protein; this prediction is based on the assumption that the linking proteins tend to have related functions then the functional prediction is calculated through the GO terms that are assigned to the neighbours which interact directly with a query protein. CONCLUSION At present time (March 2006) APID includes more than 35000 different proteins and more than 111000 protein-protein interactions, this includes interaction data that come from five main source databases (BIND, DIP, HPRD, IntAct and MINT). The three species with the larger amount of data are: Drosophila melanoganster (more than 8000 proteins and more than 31000 interactions), Homo sapiens (more than 7700 proteins and more than 21000 interactions) and Saccharomyces cerevisiae (more than 5200 proteins and more than 28000 interactions). The major overlap between databases occurs for 21000 interactions that come from 4 different databases (DIP, IntAct, MINT and BIND) but in general the overlap between databases is small. For more information APID can be explored via web (http://bioinfow.dep.usal.es/apid/) and it includes a graphic interactive tool to visualize and browse the interaction network.