Poster K-8

Dynamic query and exploration of literature paths between proteins

Authors:
Robert Hoffmann (*Memorial Sloan-Kettering Cancer Center*)
Chris Sander (*Memorial Sloan-Kettering Cancer Center*)

Short Abstract: Why do proteins X and Y show up together in a given experiment? The iHOP literature network contains the answer to many questions as to the relationships between specific proteins. However, to find these answers is difficult when considering different synonyms and allowing for indirect relationships. We present a novel literature query system that facilitates the interactive exploration of literature paths in the iHOP network.

Long Abstract:

Interpretation of experimental results forms a substantial part of an experimental biologist's day of work. Especially one question repeatedly emerges: Why do proteins X and Y show up together in a given experiment? Why do they have correlated transcription profiles, why do they co-precipitate, why co-localize in immunostaining? Answering this kind of questions always implies gathering available information on the genes involved, their cellular and physiological contexts and their specific interactions. This background knowledge can be found partially in manual curated databases[1, 2] or ontologies[3], but is primarily spread over millions of biomedical papers. With one thousand papers published every day, however, it is impossible for the individual scientist to keep up-to-date and the interpretation of novel experimental data becomes extremely elaborate. While looking up literature for N differentially regulated genes of a microarray experiment is tedious but in principal possible, the exploration of all possible relationships between these genes becomes unrealistic in practice. Approximately N2/2 PubMed queries would be necessary to explore all possible relationship between N genes and their gene products. This problem increases further, when considering all the different synonyms used for these genes in the literature.

For queries of this complexity the use of automated text-mining methods to pre-process the biomedical literature is obvious. The iHOP system, for instance, makes existing knowledge in PubMed accessible as a navigable network, were genes and proteins serve as hyperlinks between interrelated sentences and abstracts[4]. This network contains the answer to many questions as to the relationship of X and Y. However, to find these answers in a small world network like iHOP[5], is similar to the travelling salesman problem. This difficulty is even more evident, when allowing for indirect relationships between X and Y going through Z.

Here we present a novel literature query system that finds optimal graph theoretical solutions in the iHOP literature network. For instance, an experimenter can query this system with a list of genes (e.g. from a microarray) to explore all combinatorial possible sentence-pathways relating these input genes in the literature. The result of such a query could be merely numerical, however, this would neglect the textual representation, which we believe is richer in information and more intuitive to human experts. Thus we extend the well-tried iHOP user interface[5] to facilitate the exploration of found pathways within their source texts. Pathways are presented in a clickable image, where edges and nodes are hyperlinked to the

corresponding sentences that form the literature path. This way, the user can easily validate the origin and reliability of retrieved relationships. Queries can optionally be enhanced by the user through integration of interaction data from homologous organisms and large-scale experiments[6, 7]. The iHOP network is publicly accessible at http://www.ihop-net.org.

Robert Hoffmann & Chris Sander
Memorial Sloan-Kettering Cancer Center, Computational Biology Center, 1275 York Avenue, New York, NY 10021, USA.

References

1. Pruitt KD, Tatusova T, Maglott DR: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005, 33(Database issue):D501-504.
2. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R et al: The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006, 34(Database issue):D187-191.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25(1):25-29.
4. Hoffmann R, Valencia A: A gene network for navigating the literature. Nat Genet 2004, 36(7):664.
5. Hoffmann R, Valencia A: Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 2005, 21 Suppl 2:ii252-ii258.
6. Bader GD, Betel D, Hogue CW: BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 2003, 31(1):248-250.
7. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: DIP: the database of interacting proteins. Nucleic Acids Res 2000, 28(1):289-291.