Text-mining Needs and Solutions for the Biomolecular Interaction Network Database (BIND)

Ian Donaldson
Blueprint Initiative, Mount Sinai Hospital
Toronto, Ontario
Canada
ian.donaldson@mshri.on.ca

Proteomics represents a collection of experimental approaches that may be used to investigate biological systems. Such approaches commonly produce vast amounts of data that relate to physical interactions between biomolecules. One challenge in extracting useful information from these data is determining how they relate to current knowledge.

The goal of the BIND database (Biomolecular Interaction Network Database) is to curate and archive these interaction data from the literature using a standard data representation so that they may be effectively used for knowledge discovery (http://bind.ca) (Bader et al., 2001; Bader and Hogue, 2000). This database facilitates placing experimental data into context.

For instance, a biologist may be presented with evidence suggesting that several different proteins may interact with their protein of interest. One of the first obvious questions is; "Is there any evidence to support any of these potential interactions?" These questions may be answered using a number of approaches. For each of the potential interactions:

1) Are there any matching interaction records in BIND or some other interaction database?
2) Are there any interaction records that involve proteins with similar sequences?
3) Are there any interaction records that involve proteins with similar conserved domain profiles?
4) Is the potential interaction likely given the Gene Ontology annotation associated with the two interacting proteins?
5) What are the synonyms for the two potential interactors and do these synonyms ever co-occur in the literature.

Answering each of these questions requires addressing a number of technical issues, which, in principal, are trivial. However, in practice, it is non-trivial to solve all of these problems to completion and to solve them consistently. Failing to do so means that knowledge that may support a potential interaction will be lost. This is unacceptable since much of proteomics is about filtering meaningful data away from noise.

Interestingly, solving these questions is also of interest to text-miners. Mentions of any two proteins in text may be viewed as a potential interaction. A set of potential interactions may be sorted according to the answers to the above questions.

I will describe here, the ongoing efforts to incorporate the functionality of a text-mining tool called PreBIND (Donaldson et al., 2003) into a larger bioinformatics application programming platform called SeqHound. This platform already incorporates the NCBI's GenBank sequence database, Molecular Modelling Database, LocusLink and Conserved Domain database as well as functional annotation from the Gene Ontology consortium and (in the near future) interaction data from BIND (the Biomolecular Interaction Network Database). SeqHound is freely available via a web-interface and an application programming interface in C, C++, PERL and Java (http://seqhound.blueprint.org)(Michalickova et al., 2002). I envision that this system will be used by biologists to examine interaction data from high-throughput proteomics studies.

In addition, it may also be used by text-miners to help generate and submit preliminary BIND records to the BIND database.

References
Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001). BIND--The Biomolecular Interaction Network Database. Nucleic Acids Res *29*, 242-245.
Bader, G. D., and Hogue, C. W. (2000). BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics *16*, 465-477.
Donaldson, I., Martin, J., De Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K.*, et al.* (2003). PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics *4*, 11.
Michalickova, K., Bader, G. D., Dumontier, M., Lieu, H. C., Betel, D., Isserlin, R., and Hogue, C. W. (2002). SeqHound: biological sequence and structure database as a platform for bioinformatics research. BMC Bioinformatics *3*, 32.