Poster K-24

BioTermNet: a system for biomedical text mining

Authors:
Asako Koike (*Univ. of Tokyo*)
Toshihisa Takagi (*Univ. of Tokyo*)

Short Abstract: We have developed a text mining system called “BioTermNet”, which provides two functions: 1) to find implicit relationships by connecting explicit relationships for knowledge discovery, 2) to cluster concepts based on the document similarities for the interpretation of high-throughput data. These results are presented interactively by a graphic viewer.

Long Abstract:

Many experimental results have been accumulated in scientific literature as a result of rapid progress of biomedical field. Information extraction, information retrieval, and text mining techniques have become requisite for two main reasons. The first is because objective information, which is explicitly written in text, can be effectively retrieved using information extraction and information retrieval techniques. The second is because knowledge concealed within multiple sources can be discovered by connecting fragmented results described in the text. Extracting objective information is crucial to the interpretation of high-throughput analyses such as DNA/protein arrays, differential displays, and RNAi, because copious results from multiple genes can be obtained at the one time, and background knowledge buried in texts are necessary to interpret these results. The discovery of knowledge hidden in text obtained from multiple sources is advantageous in searching for new diseases to which approved drugs can be applied and in searching for what dietary effects foods have. The discovery of implicit relationships is also important in interpreting experimental results that cannot be understood from explicitly described facts in the text.

We have developed a biomedical text mining system called “BioTermNet” for knowledge discovery or hypothesis generation and interpretation of high-throughput experimental results. This system provides mainly two functions and the results are presented by a graphic viewer. The first function is to connect explicit relationships and generate the conceptual network and turn up the most appropriate implicit relationship similarly with ABC model in open and closed discovery processes. In ABC model, if the relationship between concept A and concept B is described in a paper, and that between concept B and concept C is described in another paper, the relationship between concept A and concept C can be inferred, even if there are no papers that explicitly describe the relationship between concept A and C. An “open discovery system” is where information is searched from A  B and B C, while a “closed discovery system” is where information is searched from A B and CB. In this system, multiple intermediate layers (e.g. A B B’C) are available for detailed inferences, while the conventional ABC model uses only one intermediate layer. Furthermore semantic type restriction on each layer (e.g. A B(gene product) B’(chemical compound)C (symptom)) is available for taking into account of user’s intellection. The explicit relationships are extracted from MEDLINE abstracts by syntactic analysis and distinctive co-occurrences in the same

abstract. PRIME data (http://prime.ontology.ims.u-tokyo.ac.jp) developed in our group, which contains protein-interactions, protein-diseases, and protein-functions extracted by syntactic analysis, are used in the former. The latter is calculated by Lnu term weighting[1], which was superior to other major measurements such as mutual information, Dice coefficient, and cosine coefficient in our preliminary tests. To consider the synonyms and homonyms, several dictionaries and thesauruses such as GENA (http://gena.ontology.ims.u-tokyo.ac.jp/search/servlet/gena), family name dictionary (http://gena.ontology.ims.u-tokyo.ac.jp:8081/mov/), and UMLS (Unified Medical Language System) are utilized and the gene name ambiguities are resolved as far as possible. The performance as a knowledge discovery system was demonstrated using the association between fish-oil and Raynaud's disease, that between Mg and migraine, and that between BRCA1 gene and their related diseases in the previous paper [2].

The second function of BioTermNet, which is newly developed one, is to cluster genes/concepts based on document similarities and protein interactions and provide an association matrix and its hierarchical tree for the interpretation of high-throughput experimental data such as DNA micro array and RNAi. An overview of the relationships between multiple genes/concepts can be obtained using the association matrix and its hierarchical tree, while detailed and/or causal relationships can be presented by the conceptual network or common concepts, which are obtained interactively on the graphic viewer. For example, when input data is co-expressed genes in DNA micro array, the biological-functionally related genes and/or neighboring genes on protein network are easily presented to be the same cluster in the association matrix and its hierarchical tree and bird's-eye view can be obtained, while the relationship between different functional clusters or the evidence of functional relationship between genes in the same cluster can be extracted by the conceptual network, which can be drawn interactively on the graphic viewer. The association matrices are calculated using vectors whose components are co-occurring concepts based on Lnu term weighting and protein interactions extracted by syntactic analysis, individually. The adequacies of concept clustering based on document similarities by several methods such as singular value decomposition (SVD) are compared using physically interacting gene pairs and genes with the same gene ontology annotation. When this system is used to interpret micro-array data on cancer cells, it creates a plausible distance matrix, its hierarchical tree, and an understandable conceptual network.

The BioTermNet system is available at http://btn.ontology.ims.u-tokyo.ac.jp/ for non-commercial purposes.

[1] Singhal A, Buckley C, Mitra M (1996) Pivoted Document Length Normalization. Proc of ACM SIGIR'96, 21-29.

[2] Koike A, Takagi T, Knowledge discovery based on an implicit and explicit conceptual network. J.Am. Soc. Inf. Sci. Tech., in press.