BioTermNet: a system for biomedical text mining

<u>Asako Koike</u>[1,2] , Toshihisa Takagi[1]

[1]Dept. of Computational Biology, Graduate School of Frontier Science, The University of Tokyo, Kiban-3A1 (CB01) 5-1-5, Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan

[2]Central Research Laboratory, Hitachi Ltd. 1-280 Higashi-Koigakubo, Kokubunji, Tokyo, 185-8601, Japan

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 to acquire the necessary knowledge. We have developed a biomedical text mining system called "*BioTermNet*" for knowledge discovery/hypothesis generation and interpretation of experimental results. This system provides mainly two functions and the results are presented by a graphic viewer. The first function is to connect the explicit relationships (clearly described in abstracts) and generate the conceptual network and search the most appropriate implicit relationship (not described in abstracts) for open (only start concept is given) and closed (start and end concepts are given) discoveries. The explicit relationship is calculated with a hybrid method of Lnu term weighting and protein-interactions, -diseases, and -functions using syntactic analysis (*PRIME* data, http://prime.ontology.ims.u-tokyo.ac.jp). The performance as a knowledge discovery system was validated using the association between fish-oil and Raynaud's disese and that between Mg and migraine in the previous paper [1]. The second function of *BioTermNet*, which is newly developed one, is to cluster genes/concepts based on document similarities and provide an association matrix and its hierarchical tree for the interpretation of high-throughput 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 interactively on the graphic viewer. The adequacies of concept clustering based on document similarities by several methods such as Lnu term weighting and SVD are compared using physically interacting gene pairs and genes with the same gene ontology annotation. The application of this system to the DNA array data is also discussed. The BioTermNet system is available at http://btn.ontology.ims.u-tokyo.ac.jp/ for non-commercial purposes.

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