

Additional Paper for “Biological Data Mining”

Paper reference

Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*. 2004 Oct 8;5:147.

Abstract

Biological organisms are organizations of interaction networks between genes and their environment. However, most of biological literature focuses on individual genes and interactions. Integrating such enormous amount of data is very challenging and time-consuming if done manually. Here the authors describe a natural language processing (NLP)-based text-mining algorithm, Chilobot, which constructs relationship networks based upon PubMed abstracts. It represents a new way of integrating information and visualizing biological networks. In addition, it can generate new hypothesis that may not have been directly tested yet. Finally, the authors discuss the properties derived from such networks.

Discussion

Chilobot (chip literature robot) constructs graphical representations of content-rich relationship networks between terms, which can be molecules or concepts. It does so by searching for synonyms from a number of databases (Swissprot, LocusLink GDB, HUGO, OMIM and SGD) and download abstracts that are related to these synonyms. Then Chilobot parses individual sentences and those that contains the term or its synonyms are then subjected to NLP analysis that includes softwares such as TnT (part-of-speech) and CASS (shallow parsing). Chilobot also characterizes these sentences into one of five categories: stimulatory, inhibitory, neutral, parallel and abstract co-occurrence. Then the relationships are visualized using AiSee with symbols to indicate directionality, the nature of interaction and fold-changes, etc.

Since Chilobot retrieves information from NCBI, obviously it is sensitive to records present there (91.2% successful mining), but it is insensitive to varying number of reference on the topic. In a manual validation, most of these relationships are correctly identified. In addition, in a test for “hypothetical relationship” where two terms are now known to be related but defined as “abstract co-occurrence” in earlier reference, Chilobot correctly “predicted” those interactions from those early references. This demonstrate Chilobot’s ability to form new hypothesis based on current literature. In testing 3 different networks, the distribution of the average connectivity follow a scale-free topology with the power-law ($P(k) \sim k^{-n}$, $n = 1.21$).