Poster K-25

Search Engine for Mining Biological Significance of Genetic Markers in Medline

Authors:
Weijian Xuan (*Molecular and Behavioral Neuroscience Institute and Department of Psychiatry, University of Michigan*)
Pinglang Wang (*Molecular and Behavioral Neuroscience Institute and Department of Psychiatry, University of Michigan*)
Stanley J. Watson (*Molecular and Behavioral Neuroscience Institute and Department of Psychiatry, University of Michigan*)
Fan Meng (*Molecular and Behavioral Neuroscience Institute and Department of Psychiatry, University of Michigan*)

Short Abstract: MarkerInfoFinder is the first application that allows researchers to find genetic marker related Medline records using positional and linkage disequilibrium criteria. It greatly increases the efficiency of mining literature related to a list of genetic markers, facilitating the understanding of the molecular mechanisms underlying genetic disorders.

Long Abstract:
Genome-wide high density SNP association studies are expected to identify various SNP alleles associated with different complex disorders. Understanding the pathophysiological significance of the relevant SNP alleles will be a major challenge, particularly in the context of the large body of existing literature devoted to genetic and molecular studies. This is because existing Medline search engines cannot deal with the entity name variations of different genetic markers, such as SNP, STS/microsatellite, cytoband and gene/protein allelic variants. The genomic location and the linkage disequilibrium relationships among different genetic markers are not encoded in current generation of Medline search engines either. Although extensive efforts have been devoted to the literature mining of gene and protein functions, similar work on genetic markers and their related diseases is still at its infancy.

To bridge the gap between genotyping data from genome-wide scanning and existing literature related to various forms of genetic markers, we developed a novel web-based Medline search engine, MarkerInfoFinder, for efficient literature exploration using genetic marker lists as input. It is publicly available at: http://brainarray.mhri.med.umich.edu/brainarray/datamining/MarkerInfoFinder.

Our system consists of four main function modules: (1) Identification of different types of genetic markers, including SNP, STS/microsatellite marker, cytoband and genetic variations of gene/protein, as well as disease names from Medline records; (2) Determination of positive versus negative linkage or association relationship between genetic markers and diseases; (3) Integration of marker genomic location and HapMap II data from different databases to enable the retrieval of Medline records based on user defined positional and/or linkage disequilibrium criteria; (4) A web interface that allows researchers to search Medline citations using various genetic marker IDs directly.

We incorporated a series of nature language processing techniques to identify genetic markers, gene/protein entities, and disease names in free text. For example we identified 60 patterns for cytoband and 45 patterns that are frequently used to describe gene/protein allelic variants. Corresponding rules and regular expressions were designed to capture such patterns in free text. We utilized a gene and protein identification engine developed in our group, which achieves high recall and precision and maps extracted gene/protein entities to Entrez Gene IDs. Our disease name list is based on the 2552 genetic related disorders collected in the OMIM database. We applied normalization and multiple layers of filters to the initial disease name list and identify their occurrence in Medline using a fast multikey string searching algorithm. We also implemented procedures to disambiguate extracted entities through feature term-based corpus profiling algorithm and validation through existing database records.

Since there are many papers reporting negative relationship between genetic markers and diseases, just considering co-occurrence of genetic markers and disease in individual Medline records can lead to high percentage of false positives in retrieval. We compiled 150 patterns of negation phrase and used shallow parsing and rule-based method to detect negative linkages of genetic marker and diseases. Our algorithm achieved an encouraging F-measure of 85%.

Our system integrates genetic marker data from dbSNP, UniSTS, NCBI Ideogram, and Entrez Gene. We mapped different types of marker to each other utilizing NCBI genome assembly and HapMap II data. As a result, our genetic marker search function can retrieve every type of genetic markers related to the input marker list based on flexible positional and LD criteria. On the literature side, each Medline record is indexed by every type of genetic markers and the related tables/indexes are pre-compiled. Consequently, we can easily map the query genetic marker list to related Medline citations through our genetic marker search function. Matched records will then be filtered and presented to the user. We also use MeSH descriptors to organize search results by their common themes to help researchers discover the most relevant sets of literatures. The web interface provides intuitive search functions as well as a variety of searching, displaying, sorting and exporting options for the effective exploration retrieved records.

The MarkerInfoFinder is the first application that allows researchers to find genetic marker related Medline records using positional and linkage disequilibrium criteria. It greatly increases the efficiency of mining Medline literature related to a given genetic marker or a list of genetic markers, facilitating the understanding of the molecular mechanisms underlying genetic disorders. We expect this effort will lead to the building of a knowledge base for genetic markers and various pathophysiological processes, which can be used to develop knowledge-based analysis of genome-wide scanning results.