Improving Literature Based Discovery Support by Genetic Knowledge Integration

Dimitar Hristovski[a,b], Borut Peterlin[c], Joyce A. Mitchell[b,d], Susanne M. Humphrey[b]

[a]*Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia*

[b]*Lister Hill National Center for Biomedical Communications - National Library of Medicine, Bethesda, USA*

[c]*Department of Human Genetics, Clinical Center Ljubljana, Ljubljana, Slovenia*

[d]*Department of Health Management and Informatics, School of Medicine, University of Missouri, USA*

Abstract

We present an interactive literature based biomedical discovery support system (BITOLA). The goal of the system is to discover new, potentially meaningful relations between a given starting concept of interest and other concepts, by mining the bibliographic database Medline. To make the system more suitable for disease candidate gene discovery and to decrease the number of candidate relations, we integrate background knowledge about the chromosomal location of the starting disease as well as the chromosomal location of the candidate genes from resources such as LocusLink, HUGO and OMIM. The BITOLA system can be also used as an alternative way of searching the Medline database. The system is available at http://www.mf.uni-lj.si/bitola/.

Keywords:

Medical Informatics; Knowledge Discovery; Data Mining; Medline; Discovery Support; Genes; Diseases

1. Introduction and Background

With the rapidly growing body of scientific knowledge and increasing over specialization, it is likely that the scientific work of one research group might solve an important problem that arises in the work of another group. Yet, the two groups might not be aware of the work of each other. However, a great deal of knowledge is recorded at least in a secondary form in bibliographic databases such as Medline for the field of biomedicine. Also very important for current biomedical research are various specialized molecular biology databases. In the present context, these vast databases provide both an opportunity and a need for developing advanced methods and tools for computer supported knowledge discovery.

The goal of literature-based discovery support in general is to discover new, potentially meaningful relations between a given starting concept of interest and other concepts, by mining bibliographic databases such as Medline. The idea of discovering new relations from a bibliographic database was introduced by Swanson [1] who made seven medical discoveries that have been published in relevant medical journals. The main idea is first to find all the concepts Y related to the starting concept X (e.g. if X is a disease then Y might be pathological functions, symptoms, etc.). Then all the concepts Z related to Y are found (e.g. if Y is a pathological function, Z might be a molecule, structurally or functionally, related to the pathophysiology of Y). As the last step we check whether X and Z appear together in the medical literature. If they do not appear together, we have discovered a potentially new relation between X and Z. This relation should be confirmed or rejected using human judgment, laboratory methods, or clinical investigations, depending on the

nature of X and Z.

We present an interactive biomedical discovery support system (BITOLA) for the field of biomedicine. The system can be used as a research idea generator or it can be used as an alternative method of searching Medline (first summary, then details). The intended users of the system are researchers in biomedicine. The work we present here is a continuation of our previous work in literature-based discovery support [2]. The general literature discovery algorithm described above often produces a large number of candidate relations that have to be evaluated. To decrease the number of candidate relations and to make the system more suitable for disease candidate gene discovery, we included background knowledge about the chromosomal location of the starting genetic disease as well as the chromosomal location of the candidate genes when such knowledge is available. The inclusion of this background knowledge allows us to limit the candidate genes to those that fall into the same chromosomal location as the starting disease. We extracted the disease chromosomal location from OMIM and the gene locations from the LocusLink database.

The number of genes known to be the causal agent for human diseases is growing rapidly as a by-product of the Human Genome Project. Currently a query on LocusLink will reveal almost 1350 genes that cause over 1800 diseases, giving literature references to establish the evidence by which such disease-gene associations are made. And yet there are a significant number of diseases known to be caused by genes where the exact link to a specific gene has not been made. A query of LocusLink shows that there are almost 700 genes known for the phenotype only. In these cases, the genes are known because of a specific disease phenotype, usually with a chromosomal region narrowed by linkage studies in families with the disease. These 700 genes represent diseases that are potential candidates for the techniques described in this paper and the BITOLA system.

Apart from Swanson's, related work to ours is that by Perez-Iratxeta in which they try to relate genes to genetically inherited diseases using fuzzy relations [3].

2. Materials (Databases)

The major database in our approach is *Medline,* created at the National Library of Medicine (NLM). It is the most important bibliographic database in the field of biomedicine. Each citation is associated with a set of *MeSH (Medical Subject Headings)* terms that describe the content of the item. MeSH comprises controlled vocabulary and thesaurus used for indexing articles and for searching MeSH-indexed databases, including Medline.

LocusLink [4] is developed and maintained by the National Center for Biotechnology Information (NCBI) at NLM. LocusLink provides a query interface to curated sequence and descriptive information about genetic loci in several model species including humans.

Online Mendelian Inheritance in Man (OMIM) [5] was developed and maintained at the Johns Hopkins University with the assistance of the NCBI in development for the World Wide Web and tightly integrated with LocusLink. OMIM is a catalog of human genes and genetic disorders with links to the literature.

3 Methods

3.1 Knowledge Extraction

In our system, we use Medline as the source of the known relations between biomedical concepts. We extract these relations and store them in a knowledge base. The discovery algorithm then operates on this knowledge base as described later.

In previous work, we used only the major MeSH descriptors assigned to a Medline record as a representation of the contents of the article the record is about. In the current system, being described here, we use the set of all MeSH descriptors and we analyze the full

Medline database (1966-2001). Next, we add the gene symbols that we find in the document's title and abstract fields, because in MeSH there are descriptors for very few genes. As a source of gene symbols and names, we use: the database from HGNC (HUGO Gene Nomenclature Committee), NCBI's LocusLink and OMIM. A Medline record is thus represented with the set of MeSH descriptors and gene symbols found in the title and abstract.

In an additional knowledge extraction step, we obtain the chromosomal locations for the diseases in OMIM. The gene locations are extracted from HUGO, LocusLink and OMIM.

We use association rules [6] between pairs of biomedical concepts as a knowledge extraction method with which we discover known relations between concepts. In our system an association rule of the form

$X \rightarrow Y$ (confidence, support)

means that in *confidence* percent of articles containing X, Y is also present and that there are *support* number such articles. In other words, we take concept co-occurrence as an indication of a relation between concepts. Examples of association rules include the following: *Multiple Sclerosis* → *Optic Neuritis (2.02, 117)* and *Multiple Sclerosis* → *Interferon-beta (5.17, 300)* where the concepts *Multiple Sclerosis, Optic Neuritis* and *Interferon-beta* are of semantic types *Disease or Syndrome, Disease or Syndrome* and *Pharmacologic Substance* respectively. If X is a disease, for example, then some possible relations might be: *has-symptom*, *is-caused-by*, *is-treated-with-drug* and so on. We do not extract these relations currently, but we investigate several methods of doing so. We calculated all the associations between the concepts describing a Medline record and store the calculated associations in a database management system.

3.1 Discovery Algorithm

The large association rule base is the foundation upon which the algorithm for discovering new relations between concepts works, as described in the Introduction and illustrated in Figure 1. When we are looking for a disease candidate gene, we set X to a disease of interest and Z should be of the semantic type *Gene or Gene Product.* As an additional constraint, we can require that X and Z occur at the same chromosomal location if the location information is available both for X and Z.

Our discovery support system is interactive, that is the user of the system can interactively guide the discovery process by selecting concepts and relations of interest. The system also allows the possibility of showing the Medline documents relevant to the concepts of interest as well as the related proteins and nucleotides. Because in Medline each concept can be associated with many other concepts, the possible number of $X \rightarrow Z$ combinations can be extremely large. In order to deal with this combinatorial problem, the algorithm incorporates *filtering (limiting)* and *ordering* capabilities. The related concepts can be limited by the semantic type to which they belong. We take the semantic type from the National Library of Medicine's UMLS (Unified Medical Language System). The final possibility for limiting the number of related concepts is by setting thresholds on the support and confidence measures of the association rules. The goal of the ordering is to present best candidates first to make human review as easy as possible. Currently the default ordering is by the decreasing association rule confidence, but it is also possible to order by support or semantic type.

Although our system is usable for biomedical discovery support in general, we think it is especially useful for finding new relations between diseases and genes. This is a consequence of the integration of background knowledge and is a unique feature not present in other literature based discovery support systems. There are several new possible application scenarios of our system.

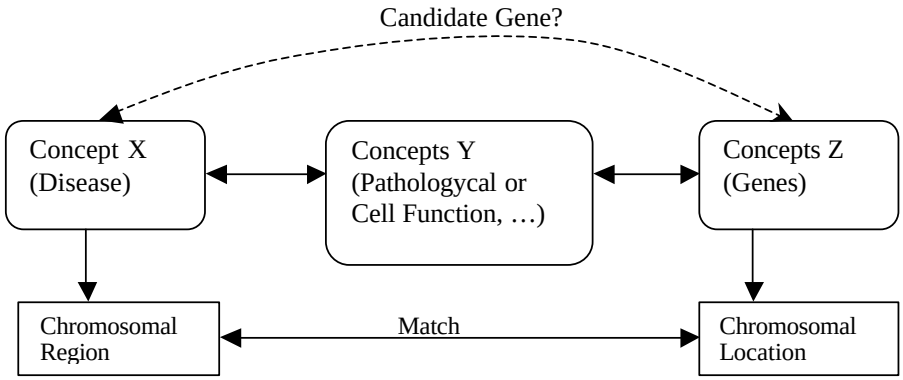

Figure 1. Discovery algorithm overview as applied to candidate gene discovery. For a starting disease X we find the related concepts Y (disease characteristics) according to the literature (Medline). Then we find the genes Z that are related to the disease characteristics Y. If the chromosomal region of the starting disease matches the location of the related genes and if there are no Medline documents mentioning both the disease X and the genes Z, then the genes Z can be proposed as candidate genes for the disease X.

In the first, we can start with a genetic disease for which the global chromosomal region is known, but not the exact gene. Through an intermediate concept (e.g. pathological or cell function) we can try to find a gene within the same region and/or expression location. The intermediate concept should reveal something about the mechanisms of the influence of the gene on the disease. In the second one, we can start with a known gene and search for a disease that might be caused or affected by that gene with a similar intermediate concept as above. In the third scenario, both the disease and gene are known to be related as a result of association or linkage study, but the nature of their relationship is not known. Here we concentrate on the intermediate concepts, which should give us an idea about the relationship.

4 Results and an Example

We analyzed the full Medline database as of the end of 2001 (11,226,520 records). We were looking for 22,252 human genes (14,659 from HUGO and 7,593 from Locus Link). To these we added 24,613 alias (synonym) gene symbols. We found a total of 10,755,807 gene symbols in 2,689,958 distinct Medline records. The most frequently found gene symbols are those that are often used with other biomedical meaning (CT, MR, CO2, …). We plan to disambiguate these gene symbols in the future. We found 29,851,448 distinct pairs of co-occurring concepts with a total co-occurrence frequency of 798,366,684 and average of 26.7. From the 29,851,448 distinct co-occurring pairs, in 7,106,099 cases at least one gene symbol appeared and in 679,159 pairs both concepts were gene symbols. From each distinct co-occurring pair (X, Y) we calculated 2 association rules (X→Y) and (Y→X) which resulted in a total of 59,702,896 association rules.

We have only done a preliminary analysis so far, but plan to evaluate the system by analyzing recently discovered disease-gene relationships and checking how many of them we are able to predict with our system using only information known prior to the publication of the relationship. This would be similar to our approach [2] and others [3].

The following is an example where the BITOLA system could be used to facilitate candidate gene examination. Polymicrogyria (PMG) is a malformation of cortical development in which the brain surface is irregular and the normal gyral pattern is replaced by multiple small, partly fused gyri separated by shallow sulci. Microscopic examination shows a simplified four-layered or unlayered cortex.

Bilateral perisylvian PMG (BPP [OMIM 260980]) often results in a typical clinical syndrome that is manifested by mild mental retardation, epilepsy, and pseudobulbar palsy, which causes difficulties with expressive speech and feeding. A locus for bilateral perysylvian polymicrogyria was recently mapped to the Xq28 region. To this region, 237 genes have been assigned as found in the Locus link database. We used BITOLA to reduce the number of candidate genes.

As the beginning concept X, we entered the name of the disease – BPP. After limiting the related concepts Y by the semantic type *Cell function*, 7 concepts were obtained: membrane potentials, cell adhesion, cell differentiation, cell movement, cell division, action potentials and phagocytosis. According to the current knowledge about development of cerebral cortex and associated malformations, we chose cell movement as the most interesting Y concept.

Using this concept, we have searched for all related concepts Z of the semantic type *Gene or Gene Product* and further limited them to those matching the chromosome location Xq28. In this manner, 18 genes were suggested by BITOLA. By using the information from the article describing localisation of BPP we excluded three genes: FMR2, GPR50 and Cxorf6 as they map centromeric to the proposed region of interest - DXS8103 – telomere.

According to the tissue specific expression which has been reviewed in Locus link, Gene Cards and OMIM we could further exclude 9 genes (ARHGAP4, F8, ATP6IP1, EMD, OPN1LW, OPN1MW, SAGE, G6PD and XM) as their expression pattern did not preferentially included brain which is obvious the target in the BPP.

Among the 6 remaining gene candidates the ABCD1 gene was associated with a rather

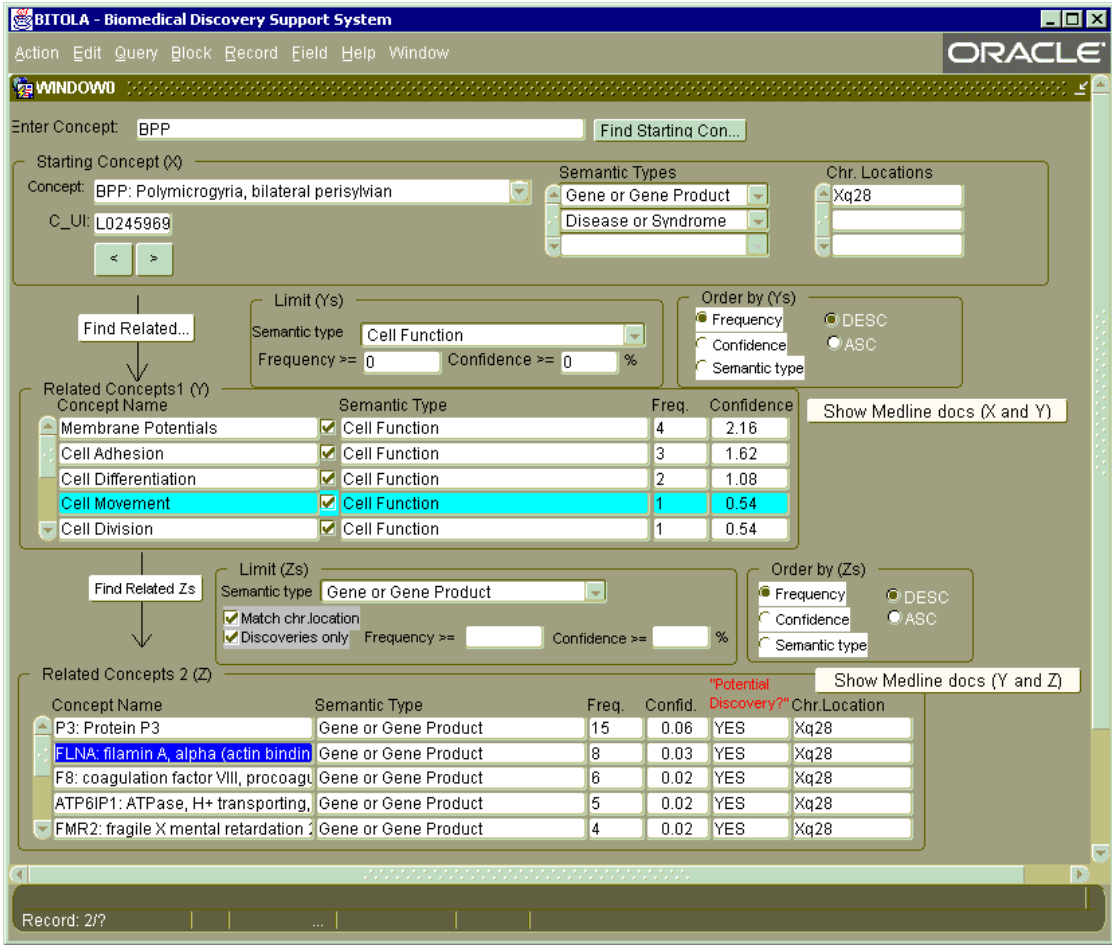

Figure 2. The BITOLA system user interface. From the starting disease *BPP: Polymicrogyria, bilateral perisylvian* through the intermediate related concept *Cell Movement*, the system offered several candidate genes for further investigation.

distinct brain disease, adrenoleucodystrophy which is mainly characterized by central nervous system demyelination. For further three genes P3, MPP1 and DXS1357E, which are expressed ubiquitously no function or disease phenotype are known so far.

Therefore two interesting candidate genes remained which could be associated with BPP: L1CAM and FLNA. L1CAM gene product is an axonal glycoprotein belonging to the immunoglobulin supergene family. This cell adhesion molecule plays an important role in nervous system development, including neuronal migration and differentiation. Mutations in the gene cause three X-linked neurological syndromes known by the acronym CRASH (corpus callosum hypoplasia, retardation, aphasia, spastic paraplegia and hydrocephalus. On the other hand, FLNA protein promotes orthogonal branching of actin filaments and links actin filaments to membrane glycoproteins. Defects in FLNA have been recently associated with periventricular heterotopia (PH). PH is an X-linked developmental dominant disorder in which many neurons fail to migrate into the cerebral cortex. Most hemizygous affected males die early during embryogenesis, whereas heterozygous females have normal intelligence but suffer from seizures and various manifestations outside the central nervous system, especially related to the vascular system. This implies that essential embryonic cell migration can only occur in FLNA- expressing cells. Since the cell function as well as several symptoms of BPP are shared to PH we believe that especially FLNA is a good candidate gene for BPP. Our hypothesis could be easily tested by mutation screening of BPP patients for mutations in FLNA gene.

5 Summary

We extend and enhance an existing interactive literature-based biomedical discovery support system (BITOLA). The system can be used as a research idea generator or as an alternative method of searching Medline. To decrease the number of candidate relations and to make the system more suitable for disease candidate gene discovery, we include genetic knowledge about the chromosomal location of the starting disease as well as the chromosomal location of the candidate genes.

References

[1] Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986 Autumn;30(1):7-18.

[2] Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Medinfo. 2001;10(Pt 2):1344-8.

[3] Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nat Genet. 2002 Jul;31(3):316-9.

[4] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001 Jan 1;29(1):137-140.

[5] Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2002 Jan 1;30(1):52-5.

[6] Agrawal, R. et al: Fast discovery of association rules. In U. Fayyad et al, editors, Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA. (1996)

Address for correspondence

Dimitar Hristovski. Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana; Vrazov trg 2/2 1104 Ljubljana, Slovenia. E-mail: Dimitar.Hristovski@mf.uni-lj.si