International Journal of Medical Informatics (2005) **74**, 289—298

ELSEVIER

International Journal of Medical Informatics

www.intl.elsevierhealth.com/journals/ijmi

Using literature-based discovery to identify disease candidate genes

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, Vrazov trg 2/2 1104 Ljubljana, Slovenia*
[b] *Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, USA*
[c] *Division of Medical Genetics, University Medical Centre Ljubljana, Ljubljana, Slovenia*
[d] *Department of Health Management and Informatics, School of Medicine, University of Missouri, USA*

Received 25 November 2003; received in revised form 6 April 2004; accepted 20 April 2004

KEYWORDS
Medical informatics;
Knowledge discovery;
Data mining;
MEDLINE;
Discovery support;
Genes;
Diseases

Summary We present BITOLA, an interactive literature-based biomedical discovery support system. The goal of this 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 and Human Genome Organization (HUGO). BITOLA can also be used as an alternative way of searching the MEDLINE database. The system is available at http://www.mf.uni-lj.si/bitola/.

1. Introduction

With the rapidly growing body of scientific knowledge and increasing specialization, it is possible that the research of one group might solve an important problem of another, without the two groups being aware of each other's work. The goal of literature-based discovery is to address this situation by uncovering new, potentially meaningful relations between a starting concept of interest and other concepts. In the field of biomedicine, a great deal of knowledge is recorded at least in secondary form in bibliographic databases such as MEDLINE as well as various specialized molecular biology databases. These resources provide both an opportunity and a need for developing advanced methods and tools for computer-supported knowledge discovery. For example, we might devise a system that looks for genes that cause a particular disease or for drugs that treat that disease.

* Corresponding author.
E-mail address: dimitar.hristovski@mf.uni-lj.si (D. Hristovski).

1386-5056/$ — see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2004.04.024

In this paper we present an interactive biomedical discovery support system (called BITOLA) for the field of biomedicine, particularly for discovering candidate genes in etiological relationships with diseases. 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 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 to LocusLink reveals over 1600 genes that cause over 2000 diseases, giving literature references to establish the evidence by which such disease—gene associations are made. 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 to LocusLink shows over 765 genes known forphenotype 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 765 genes represent diseases that are potential candidates for the techniques described in this paper and the BITOLA system.

2. Background

The idea of discovering new relations from a bibliographic database was introduced by Swanson [1] who, together with Smalheiser, 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.

Gordon and Lindsay [2] and Weeber et al. [3] have repeated some of the Swanson's discoveries with different methods. Weeber et al. has also discovered several hypothetical new therapeutic applications of existing drugs [4]. Recently, Srinivasan [5] also replicated most of the Swanson and Smalheiser's discoveries with a system based on concept profiles consisting of weighted Medical Subject Headings (MeSH®) terms. However, they do not explicitly deal with genetic knowledge.

The intended users of the BITOLA system are researchers in biomedicine. The work we present here is a continuation of our previous work in literature-based discovery support [6]. 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 chromosomal locations and regions from the LocusLink database.

Related to our work is the information extraction research in gene—gene interactions and relations. Jenssen et al. [7] build a literature network of human genes. Stapley and Benoit [8] build network of co-occurring genes from a small MEDLINE subset for information retrieval and visualization. Also concerned with discovering gene relations and interactions is the work of Stephens et al. [9] and Sekimizu et al. [10]. Rindflesch [11] uses a natural language processing method for extracting causal relations between genetic phenomena and diseases. In these approaches knowledge discovery is not explicitly the goal, but it is expected to happen by visualization of the relations by the user.

Related research in the field of disease candidate gene prediction is that published in [12—15]. In [12], they try to relate genes to genetically inherited diseases using fuzzy relations. In [13], they predict disease relevant human genes by clustering diseases based on their phenotypic similarity. In [14], for given positional and expression/phenotypic data, they integrate relevant data from several database to produce a quick overview of interesting genes. In [15], they present a computational approach to prioritize candidate disease genes that is based on over-representation of functional annotation between loci for the same disease.

3. Materials

The major database in our approach is MEDLINE, created at the National Library of Medicine® (NLM®). Each citation is associated with a set of

MeSH terms that describe the content of the associated article. MeSH is a controlled vocabulary and thesaurus used for indexing articles and for searching MeSH-indexed databases, in particular, MEDLINE.

The Unified Medical Language System® (UMLS®) project that NLM began in 1986 was undertaken in order to provide a mechanism for linking diverse medical vocabularies as well as sources of information. UMLS currently contains three components: the Metathesaurus®, Semantic Network, and SPECIALIST Lexicon [16,17]. The Semantic Network contains information about the types or categories (e.g., "Disease or Syndrome", "Virus") to which all concepts in the Metathesaurus have been assigned.

LocusLink [18] 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. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, Enzyme Commission (EC) numbers, Mendelian Inheritance in Man (MIM) numbers, UniGene clusters, homology, map locations, and related websites. It provides direct links to the PubMed® literature as well as links to other gene centered resources.

HUGO has as one branch of its mission to promote international discussion and collaboration on scientific issues and topics crucial to the progress of the world-wide human genome initiative in order that the analysis of the human genome can be achieved as rapidly and effectively as possible. As such they have a Human Genome Nomenclature Committee (HGNC) responsible for the approval of a unique symbol for each gene, and also designating a longer and more descriptive name [19].

4. Methods

4.1. System overview

In Fig. 1 we can see a conceptual overview of the BITOLA system and its components. From a set of input databases, using a knowledge extraction process, we build the knowledge base of the system. The input databases represent the known biomedical knowledge. We represent this knowledge in the knowledge base in a formal form as a set of biomedical concepts, association rules between these concepts and additional background knowledge about the concepts. The association rules (based on concept co-occurrence and described in section "Knowledge extraction") are used to represent the known relations between the concepts. The discovery algorithm proposes new relations between the concepts based on the association rules (known relations) and taking into consideration the background knowledge. The user of the system, usually a researcher, interactively controls and guides the discovery process through a user interface. Now we describe in more detail the basic components of the system.

4.2. Knowledge extraction

BITOLA uses MEDLINE as the source of the known relations between biomedical concepts. In previous work, we used only the major (central concept)

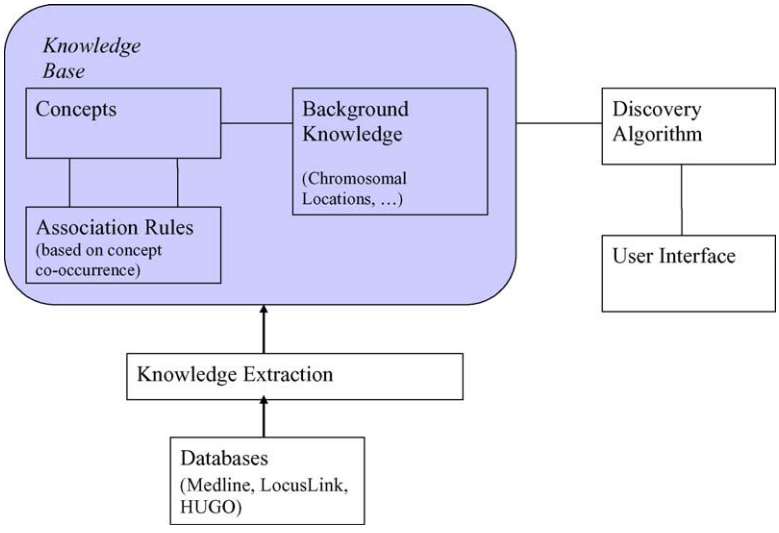

Fig. 1 BITOLA system overview.

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 all the assigned 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 to make up for the fact that MeSH has few descriptors for genes. As a source of gene symbols and names, we use: the database from HGNC (HUGO Gene Nomenclature Committee) and NCBI's LocusLink. One gene can have more then one gene symbol, and in a considerable number of cases one gene symbol can be a symbol for more then one gene. When we find one of these genes, we count as if we have found all of the genes that share the same gene symbol. A MEDLINE record is thus represented with the set of MeSH descriptors and gene symbols found in the title and abstract. Thus, the set of concepts we operate on consists of the union of the MeSH descriptors and genes from LocusLink and HUGO. We organize the set of genes similarly to the UMLS-each gene represents one concept and for each gene we have different names and symbols. Additionally, we introduced two new semantic types to classify the genes. One is *Gene or Gene Product*, which represents the genes with known sequence and product, and the other semantic type is *Genetic Disease*, which represents the genes with LocusLink type *Phenotype Only*.

In an additional knowledge extraction step, we obtain the chromosomal locations for the diseases and genes from HUGO and LocusLink.

We use association rules [20] between pairs of biomedical concepts as a method to discover known relations between concepts according to the medical literature. Our use of association rules is unconventional in the sense that we use them to discover known relations and not unknown. Association rules were originally developed with the purpose of market-basket analysis, where it is of interest to find patterns of the form $X \rightarrow Y$, with the intuitive meaning "baskets that contain X tend to contain Y". In our system an association rule has the form

$$X \rightarrow Y \text{ (confidence, support)}$$

where *support* (or frequency of co-occurrence) is the number of records with X and Y in common, and *confidence* is the percentage of records containing Y within all records containing X. In other words, we take concept co-occurrence in MEDLINE as an indication of a relation between concepts. We calculated a priori all the associations between the concepts describing a MEDLINE record and stored the calculated associations in a database management system. Examples of association rules include the following: *Multiple Sclerosis* $\rightarrow$ *Optic Neuritis* (2.02, 117) and *Multiple Sclerosis* $\rightarrow$ *Interferon-beta* (5.17, 300) where support for the former is 117 documents at confidence level of 2.02%, and for the latter, 300 documents at confidence 5.17%. 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 are investigating several methods of doing so. However, if requested by the user, we do provide a way to show MEDLINE records that support the association rule, i.e. MEDLINE records in which the two concepts forming the association rule co-occur. The BITOLA system builds a search request, starts an external web browser, connects to the PubMed website and executes the prepared search request. By reading the resulting MEDLINE records, the user should get an insight about the relation between X and Y in the association rule $X \rightarrow Y$.

4.3. Discovery algorithm

The large association rule base is the foundation for the algorithm discovering new relations between concepts as described in Table 1 and illustrated in Fig. 2. For a given starting concept X (e.g. a disease), first we find all the related concepts Y (e.g. pathological functions, symptoms, etc.). Then we find the concepts Z related to Y (e.g. if we are looking for a disease candidate gene, Z should be of the semantic type Gene or Gene Product). Next, we eliminate those concepts Z whose chromosomal location does not match the chromosomal region of the starting concept X. Of course, the chromosomal location matching is optional and applicable only if

Table 1 The algorithm for discovering new relations between medical concepts

1	Let X be a given starting concept of interest
2	Find all concepts Y such that there is an association rule $X \rightarrow Y$
3	Find all concepts Z such that there is an association rule $Y \rightarrow Z$
4	Eliminate those Z whose chromosomal location does not match the location of the starting concept X
5	Eliminate those Z for which an association $X \rightarrow Z$ already exists
6	The remaining Z concepts are candidates for a new relation between X and Z
7	Rank and display the remaining Z concepts

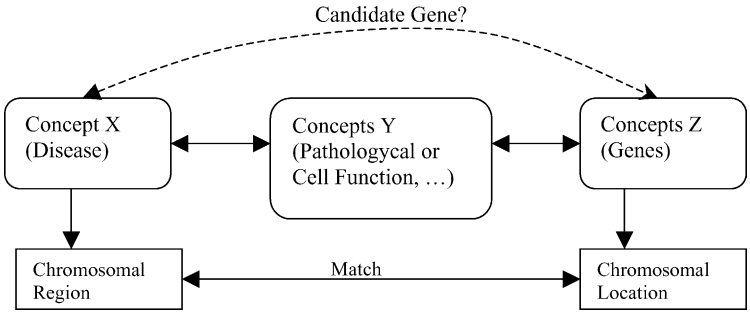

Fig. 2 Discovery algorithm overview as applied to candidate gene discovery. For a starting disease X we find the related concepts Y (disease characteristics) according to the literature (MEDLINE). Then we find the genes Z that are related to the disease characteristics Y. If the chromosomal region of 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.

the locations of the concepts X and Z are both available. For the remaining concepts Z, we check if 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. The remaining concepts Z are now ranked and displayed to the user for evaluation and further investigation. It should be stressed that, in general, it is possible to have more then one intermediate concept Y on the path from X to Z, and it is also possible to get from X to Z through different paths (Fig. 3).

For ranking the related concepts Z we use a heuristic ranking function, which takes into account that we can get from X to Z through several intermediate concepts Y (Fig. 3). The ranking function, shown in Eq. (1), is constructed in such a way to give higher rank if there are more paths from X to Z and if both relations X → Y and Y → Z are strong. Here, the calculation of rank is based on the support, but it can be calculated based on the confidence as well. In this equation, Z_k is the concept whose rank is calculated, S_{XY_i} and $S_{Y_iZ_k}$ are the supports of the association rules X → Y_i and Y_i → Z_k, respectively, and m is the number of intermediate concepts Y.

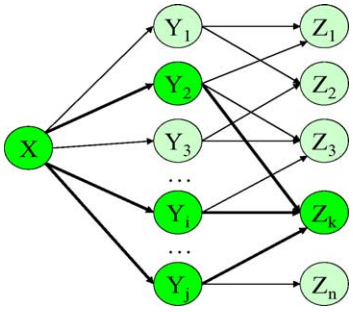

Fig. 3 Ranking the Z concepts considering multiple paths from the starting concept X through several intermediate Y concepts.

$$\text{Rank}_S(Z_k) = \sum_{i=1}^{m} (S_{XY_i} \times S_{Y_iZ_k}) \quad (1)$$

Because in MEDLINE each X concept can be associated with many Y concepts, each of which can be associated with many Z concepts, the possible number of X → Z combinations can be extremely large. In order to deal with this combinatorial problem, the algorithm incorporates filtering (limiting) and ordering capabilities. By filtering, we try to limit the number of X → Y or Y → Z associations and to minimize the number of accidental associations. The filtering possibilities are optional and can be interactively enforced by the user of the system. The semantic type to which they belong can limit the related concepts. The information about the semantic types to which a concept belongs is drawn from the Semantic Network component of the UMLS. The last possibility for limiting the number of related concepts is by setting thresholds on the support and confidence measures of the association rules in steps 2 and 3 of the algorithm.

To make human review as easy as possible, the system provides concept ordering. The related concepts Y can be ordered by association rule support (co-occurrence frequency), confidence or semantic type. The related concepts Z can be ordered by their support or confidence rank as described in Eq. (1), the default being by descending support rank.

5. Results and an example

Although BITOLA can be used 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.

In one scenario, we might 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 a second possible application, we 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.

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 7593 from LocusLink). 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, etc.). We plan to disambiguate these gene symbols in the future using JD indexing as described in the next section. 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 two association rules ($X \rightarrow Y$) and ($Y \rightarrow X$) 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 [6] and others [12]. The following is an example where BITOLA 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 perisylvian polymicrogyria was recently mapped to the Xq28 region. To this region, 237 genes have been assigned as found in the LocusLink database. We used BITOLA to reduce the number of candidate genes (Fig. 4).

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*, seven concepts were obtained corresponding to MeSH descriptors: 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 LocusLink, Gene Cards and OMIM we could give lower likelihood to nine genes (ARHGAP4, F8, ATP6IP1, EMD, OPN1LW, OPN1MW, SAGE, G6PD and XM) as their expression pattern did not preferentially included the brain which is obvious the target in the BPP.

Among the six remaining gene candidates the ABCD1 gene was associated with a rather distinct brain disease, adrenoleukodystrophy, and so it is less likely to be a candidate gene although we do not exclude it. Adrenoleukodystrophy is an X-linked disorder and results in the apparent defect in peroxisomal beta oxidation and the accumulation of the saturated very long chain fatty acids in all tissues of the body. The manifestations of the disorder occur primarily in the adrenal cortex, the myelin of the central nervous system, and the Leydig cells of the testes. For further three genes P3, MPP1 and DXS1357E, which are expressed ubiquitously no function or disease phenotype is known so far.

Therefore, two interesting candidate genes remained which could be associated with BPP: L1CAM

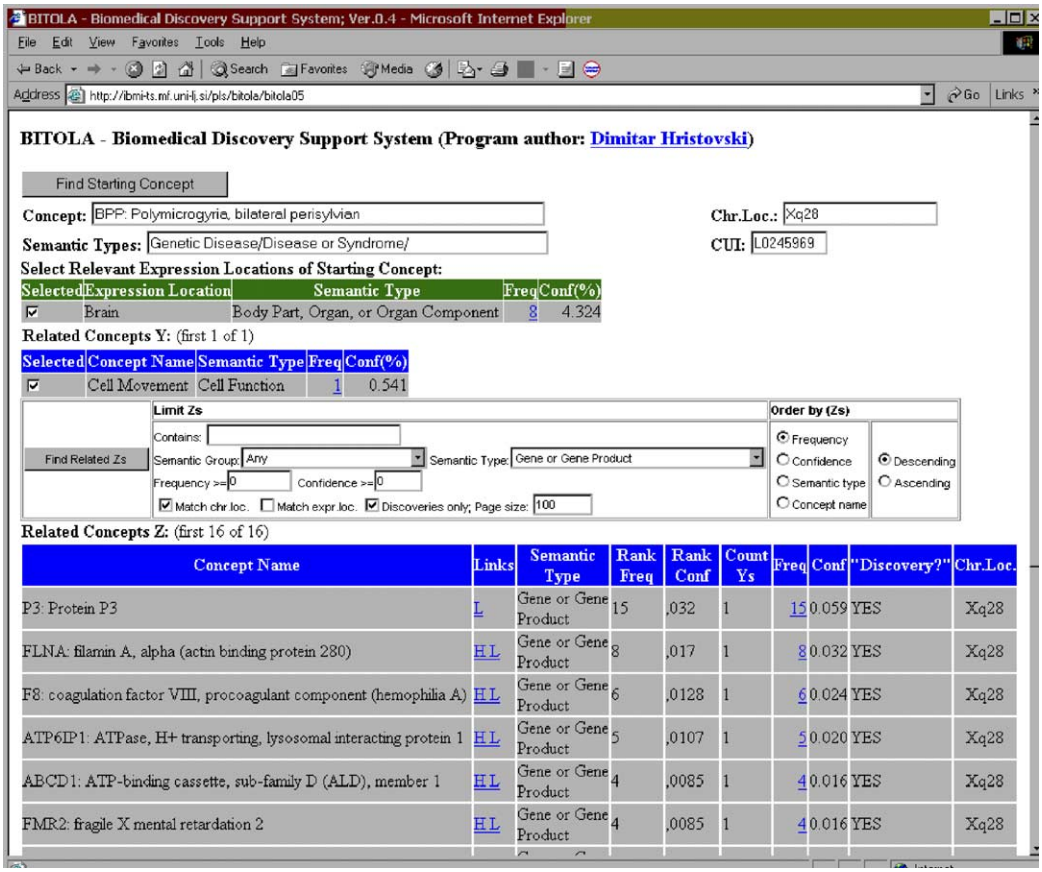

Fig. 4 The BITOLA system user interface. From the starting disease BPP: polymicrogyria, bilateral perisylvian through the intermediate related concept cell movement, the system offered several candidate genes for further investigation.

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, spasticparaplegia and hydrocephalus). On the other hand, FLNA protein promotes orthogonal branching of actin filaments and links actin filaments to membrane glycoproteins. Defects in FLNA recently have been 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 readily tested by mutation screening of BPP patients for mutations in FLNA gene.

6. Discussion

We would like to highlight in this section some of the terminology problems we faced during the development of BITOLA. MeSH is the primary source of concepts in our literature-based discovery approach. However, MeSH does not contain many specific genetic diseases and is struggling to add specific MeSH terms for the human genes. Consequently, we were forced to detect and extract the gene symbols from the title and abstract MEDLINE fields. We also propose a solution for one of the major problems with gene symbol detection, namely gene symbol disambiguation.

6.1. MeSH

One difficulty with the study of genetic diseases via the literature is that the MeSH indexing terms

often are not specific enough to allow retrieval on specific disease subtypes, and MEDLINE users must then rely on text word searching. In an evaluation of UMLS as a knowledge resource for biomedical informatics, we considered a set of 1700 diseases that were specific OMIM phenotypes that had been associated with a specific gene. Only 34% of those diseases were specifically matched to concepts in the UMLS [21]; not all of these UMLS concept matches were from MeSH. The result is a dearth of MeSH indexing terms related to specific genetic diseases, especially when the diseases include subtypes. One example demonstrating the difficulty with MeSH terms and genetic diseases is Limb Girdle Muscular Dystrophy (LGMD). LGMD is only one of many types of muscular dystrophy; LGMD has been separated into 13 separate disease subtypes that are caused by 13 separate genes giving rise to both autosomal dominant and autosomal recessive inheritance patterns. But the MeSH heading encompasses all muscular dystrophies and therefore does not facilitate such fine-grained distinctions.

6.2. Gene names and gene symbols

The gene names and gene product names are especially complex. The number of synonyms and the non-intuitive nature of the synonyms for various diseases, genes, and gene symbols make it difficult to find comprehensive information, and there is no one place to find all of the synonyms [22]. Using MeSH to go from either gene names or gene symbols to find the related literature is as problematic as the path from genetic diseases to the literature. We analyzed the loci that were part of the NLM's Gene Indexing project to link the literature to specific LocusLink records [23]. This study demonstrated that 45% of human genes had a specific MeSH term with another 37% having a supplementary concept record (SCR), another type of specific entry term. However, this leaves 17% of loci with no specific entry term. And most searchers would not have knowledge of the existence of the SCR's and so find it difficult to use them as search aids.

Many of the SCR's are gene symbols. Gene symbols are short acronyms that often create ambiguities if used outside the context of gene names. For example, the BACH (brain acyl-CoA hydrolase) protein is seen as the name of a famous composer. This contributes to difficulties with searching for information about these proteins in sources outside of molecular biology focused information resources. The ambiguity, even in biomedical sources, leads to the problem of falsely identifying gene symbols or names in documents. The overlap with standard medical acronyms and abbreviations is evident with such gene symbols as CO_2 (complement component 2) and COPD (coatomer protein complex, subunit delta). Further ambiguity arises when the gene symbol is identical to the disease name such as is seen in the CF gene (CFTR) that causes the CF disease (cystic fibrosis).

6.3. Gene symbol disambiguation

Finally, to help with disambiguating gene symbols in the millions of documents in MEDLINE, we are considering methods designed to characterize documents very broadly according to discipline, and therefore having potential to filter out non-relevant (i.e., non-genetics) documents early on. For this, we intend to explore journal descriptor (JD) indexing, developed by Humphrey [24], which uses the fact that NLM indexes MEDLINE journals per se using a small set of 127 MeSH descriptors corresponding to biomedical disciplines. For example, the journal Human Molecular Genetics is described by two JDs: Cytogenetics and Genetics, Medical. This means that each document from this journal, in effect, inherits these two descriptors as very broad MeSH indexing terms for the document. From a training set of 435,000 MEDLINE documents indexed in 1999, co-occurrence data have been extracted between the inherited JDs and MeSH headings (MHs) that had been assigned to the documents by human indexers.

For each MH in the training set, the system has computed, for each of the JDs, the ratio between the number of co-occurrences of the MH with the JD, and the number of documents indexed with the MH. For example, for the MH DNA and the JD Cytogenetics, this ratio would be 264 (co-occurrences of DNA and Cytogenetics) divided by 5361 (document count for DNA) = 0.0492, which is the JD ranking of Cytogenetics for DNA. Some JDs, for example Sociology, have 0 co-occurrences with DNA. The 127 JD rankings for DNA, ordered alphabetically by JD, form a JD vector.

For each MH in a document (outside the training set) that we wish to index broadly, the system uses its JD vector from the training set. A new JD vector, which is the centroid of these JD vectors for the assigned MHs, is then computed, consisting of the average of the rankings for each JD across the MHs. This new vector, when ordered by rank, gives a numerical association, ranging from 0 to 1, between the document and each JD. The most highly ranked JDs can be assumed to be the best JD indexing terms for the document. We suppose that the presence of the following JDs among the top-ranked JDs

for a document ought to be indicative of genetics-related content: Biotechnology; Cytogenetics; Genetics, Biochemical; Genetics, Medical and Molecular Biology. JD indexing has been enhanced to generate further general indexing according to the 127 of the 135 semantic types (STs) developed by NLM [25]. STs related to the field of genetics include Gene or Genome; Genetic Function; Molecular Biology Research Technique; Nucleic Acid Sequence; and Nucleic Acid, Nucleoside, or Nucleotide.

In preliminary experiments, about 2100 documents identified as containing gene symbols by BITOLA underwent JD and ST indexing. Based on JD and ST ranking criteria, a document determined to be in the field of genetics served to confirm the identified symbol as indeed being a gene symbol; otherwise, it is being a gene symbol was refuted. For example, JD indexing for a document titled "A yeast model for classical juvenile Batten disease (CLN3)" met the threshold for a genetics document, thus supporting CLN3 as a gene symbol. In contrast, a document titled "Ethics in a twist: 'Life Support', BBC1" did not meet this threshold, thereby differentiating the BBC1 in this title (which was a British television station featuring a new drama series called "Life Support") from the BBC1 gene symbol for "breast basic conserved 1 gene". Continuing this exploratory research, we hope to be able to demonstrate that JD and ST indexing can help to reduce the time and effort needed to identify genetic information from large document collections.

7. Conclusion

We present an 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.

Acknowledgments

Part of this research was conducted while Dimitar Hristovski was a postdoctoral fellow at the National Library of Medicine (NLM). Dimitar Hristovski thanks NLM and the ORISE program for support. In addition, we would like to thank Tom Rindflesch and Neil Smalheiser for providing valuable comments and insights.

References

[1] D.R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, Perspect. Biol. Med. 30 (1) (1986) 7—18.
[2] M.D. Gordon, R.K. Lindsay, Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil, J. Am. Soc. Inf. Sci. 47 (2) (1996) 116—128.
[3] M. Weeber, H. Klein, L.T.W. De Jong-Van Den Berg, R. Vos, Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries, J. Am. Soc. Inf. Sci. Technol. 52 (7) (2001) 548—557.
[4] M. Weeber, R. Vos, H. Klein, L.T.W. De Jong-Van Den Berg, A.R. Aronson, G. Molema, Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide, J. Am. Med. Inform. Assoc. 10 (3) (2003) 252—259.
[5] P. Srinivasan, Text mining: generating hypotheses from MEDLINE, J. Am. Soc. Inf. Sci. Technol. 55 (5) (2004) 396—413.
[6] D. Hristovski, J. Stare, B. Peterlin, S. Dzeroski, Supporting discovery in medicine by association rule mining in MEDLINE and UMLS, Medinformation 10 (2) (2001) 1344—1348.
[7] T.K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, A literature network of human genes for high-throughput analysis of gene expression, Nat. Genet. 28 (1) (2001) 21—28.
[8] B.J. Stapley, G. Benoit, Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts, Pac. Symp. Biocomput. (2000) 529—540.
[9] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, J. Mostafa, Detecting gene relations from MEDLINE abstracts, Pac. Symp. Biocomput. (2001) 483—495.
[10] T. Sekimizu, H.S. Park, J. Tsujii, Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts, in: Genome Inform Ser. Workshop Genome Inform., vol. 9, 1998, pp. 62—71.
[11] T.C. Rindflesch, B. Libbus, D. Hristovski, A.R. Aronson, H. Kilicoglu, Semantic relations asserting the etiology of genetic diseases, in: Proceedings of the AMIA Symposium, 2003, pp. 554—558.
[12] C. Perez-Iratxeta, P. Bork, M.A. Andrade, Association of genes to genetically inherited diseases using data mining, Nat. Genet. 31 (3) (2002) 316—319.
[13] J. Freudenberg, P. Propping, A similarity-based method for genome-wide prediction of disease-relevant human genes, Bioinformatics 18 (Suppl. 2) (2002) S110—S115.
[14] M.A. van Driel, K. Cuelenaere, P.P. Kemmeren, J.A. Leunissen, H.G. Brunner, A new web-based data mining tool for the identification of candidate genes for human genetic disorders, Eur. J. Hum. Genet. 11 (1) (2003) 57—63.
[15] F.S. Turner, D.R. Clutterbuck, C.A. Semple, POCUS: mining genomic sequence annotation to predict disease genes, Genome Biol. 4 (11) (2003) R75.
[16] Unified Medical Language System (UMLS), January 28, 2003. http://www.nlm.nih.gov/research/umls/.
[17] B.L. Humphreys, D.A.B. Lindberg, H.M. Schoolman, G.O. Barnett, The Unified Medical Language System: an informatics research collaboration, JAMIA 5 (1) (1998) 1—11.
[18] K.D. Pruitt, D.R. Maglott, RefSeq and LocusLink: NCBI gene-centered resources, Nucl. Acids Res. 29 (1) (2001) 137—140.

[19] H.M. Wain, R.C. Lovering, E.A. Bruford, M.J. Lush, M.W. Wright, S. Povey, Guidelines for human gene nomenclature, Genomics 79 (4) (2002) 464—470.
[20] R. Agrawal, et al., Fast discovery of association rules, in: U. Fayyad, et al. (Eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1996.
[21] O. Bodenreider, J.A. Mitchell, A.T. McCray, Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics, in: Proceedings of the AMIA Symposium, 2002, pp. 61—65.
[22] J.A. Mitchell, A.T. McCray, O. Bodenreider, From phenotype to genotype: issues in navigating the available information resources, Meth. Inf. Med. 42 (2003) 557—563.
[23] J.A. Mitchell, A.R. Aronson, J.G. Mork, L.C. Folk, S.M. Humphrey, J.M. Ward, Gene indexing: characterization and analysis of NLM's GeneRIFs, J. AMIA, Suppl. (2003) 460—465.
[24] S.M. Humphrey, Automatic indexing of documents from journal descriptors: a preliminary investigation, J. Am. Soc. Inf. Sci. 50 (8) (1999) 661—674.
[25] S.M. Humphrey, T.C. Rindflesch, A.R. Aronson, Automatic indexing by discipline and high-level categories: methodology and potential applications, in: D. Soergel, P. Srinivasan, B. Kwasnik (Eds.), Proceedings of the 11th ASIST SIG/CR Classification Research Workshop, November 12, 2000, Chicago, American Society for Information Science and Technology, Silver Spring, MD, 2000, pp. 103—116.