Literature Based Discovery Support System and its Application to Disease Gene Identification

Dimitar Hristovski[1], Borut Peterlin[2], Sašo Džeroski[3], Janez Stare[1]

[1]IBMI, Medical Faculty; Vrazov trg 2/2, 1105 Ljubljana, Slovenia
dimitar.hristovski@mf.uni-lj.si
janez.stare@mf.uni-lj.si
[2]Department of Human Genetics, Clinical Center Ljubljana; Zaloska, 1000 Ljubljana, Slovenia
borut.peterlin@guest.arnes.si
[3]Institute Jozef Stefan; Jamova 39, 1000 Ljubljana, Slovenia
saso.dzeroski@ijs.si

Abstract. We present an interactive discovery support system, which for a given starting concept of interest, discovers new, potentially meaningful relations with other concepts that have not been published in the medical literature before. The known relations between the medical concepts come from the MEDLINE bibliographic database and the UMLS (Unified Medical Language System). We use association rules to mine for the relationships between medical concepts. Then we demonstrate a successful application of the system for predicting a gene candidate for a disease, a fact recently confirmed via the positional cloning approach. We conclude that the discovery support system we developed is a useful tool complementary to the already existing bioinformatic tools in the field of human genetics. The system described in this chapter is available at http://www.mf.uni-lj.si/bitola/

1. Introduction

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 main points addressed in this paper are: 1) Is it possible to discover new, potentially meaningful relations (knowledge) between medical concepts by searching and analyzing the documents from a bibliographic database such as MEDLINE? 2) To what degree can the discovery process be automated? and 3) Can this process be used for candidate gene discovery for a human disease? To deal with these issues, we developed an interactive discovery support system based on association rule mining of the MEDLINE bibliographic database. Its intended use is as a generator for research ideas that should be then investigated by traditional scientific methods.

2. Background

The idea of using a bibliographic database for generating new medical discoveries that need to be later verified by traditional follow-up studies was proposed by Swanson [1]. He managed to make seven medical discoveries just by searching the MEDLINE database with some smart strategies and by analyzing the resulting bibliographic records. These discoveries were later confirmed and published in relevant medical journals. Swanson's discovery support process is based on the concepts of complementary literatures and noninteractive literatures. If one set of articles (XY) reports an interesting relation between concepts X and Y, and a different set of articles (YZ) reports a relation between Y and Z, but nothing has been published concerning a possible link between X and Z, then XY and YZ are called complementary literatures. Generally, XY and YZ are complementary if a potentially new relation can be inferred by considering them together that cannot be inferred from either of them separately. For example, X might be a disease, Y a physiological function associated with X, and Z a substance or drug which induces or regulates the physiological function Y. If the readers and authors of one body of literature are not acquainted with another, as might often be the case with two different specialties, then the two literatures are noninteractive. By combining the concepts of complementary and noninteractive literatures, Swanson developed the concept of undiscovered public knowledge, meaning that although the literatures XY and YZ represent publicly available knowledge, the potentially new relation between X and Z remains undiscovered and is a valuable source of new discoveries.

The first published example of a discovery Swanson made was about Raynaud's disease and fish oil [1]. Articles on Raynaud's disease (X) and articles on eicosapentaenoic acid (Z) when considered together indicated that dietary fish oil rich in eicosapentaenoic acid might be beneficial for treating Raynaud patients. One Y concept was, for example, blood viscosity. The line of reasoning used was: dietary eicosapentaenoic acid (Z) can decrease blood viscosity (Y), which has been reported in patients with Raynaud's disease (X).

Another notable discovery made by Swanson was about the relation between migraine (X) and magnesium deficiency (Z) [2]. It was discovered that magnesium deficiency was the cause of certain physiological effects (Y), which were associated with migraine.

In the beginning, Swanson performed the discovery process manually by searching the MEDLINE database. Later he added software support for some of the stages of the process. His current system is called ARROWSMITH and is described in detail in [3]. Swanson's discovery methodology contains two steps that are usually done sequentially, but each one can be done independently as well.

The goal of the first step is, for a given starting concept X, to find potentially new relations to concepts Z that are unknown at the beginning. The user starts by searching MEDLINE for all the articles about a starting concept of interest (X). The articles found are then uploaded into ARROWSMITH. Then the titles of the articles are analyzed and a list of all words and phrases is made. This list is, of course, very

large, and a stop word list is used to reduce it. Another way to alter the list is by manual user editing. The remaining words and phrases are considered concepts (Y) that are somehow related to the starting concept (X). Now a set of search strategies is generated in order to search MEDLINE for each of the Y concepts. The search results are uploaded into the system and again a list of words and phrases appearing in the titles is produced. This is the set of concepts Z related to Y. Those Z concepts for which there are articles in MEDLINE containing both X and Z are eliminated. The remaining concepts in Z represent possible candidates of novel relations between X, Y and Z, where Y is some intermediary concept linking X to Z.

The goal of the second step of the Swanson's discovery methodology is, for given concepts X and Z, to find intermediate concepts Y through which X and Z are related. It is possible that more than one Y leads from X to Z and ARROWSMITH orders the Z concepts by decreasing number of Y connections. So, in ARROWSMITH, the frequency of words or phrases in article titles is used as a measure of relational strength between medical concepts. The publicly available version of ARROWSMITH supports only the second step of the discovery methodology.

Some of the Swanson's discoveries were repeated with different methods by Gordon and Lindsay [4], and by Weeber [5]. Weeber also discovered several hypothetical new therapeutical applications of existing drugs. For more details about the Swanson's approach and its comparison with the others, see the Weeber chapter in this volume.

Our system is based on Swanson's ideas, but there are however, several notable differences between our approach and theirs. Instead of using title words as a representation of the meaning of the MEDLINE documents, we use MeSH (Medical Subject Headnings) descriptors. We use association rules as a measure of relationship between medical concepts while Swanson uses word frequencies. We have built a large association rule base by pre-calculating and storing the association rules in a database management system. This allows us to build a truly interactive discovery support system with fast response.

3. Materials and Methods

MEDLINE

The MEDLINE database is a product of the US National Library of Medicine (NLM). Because of its coverage and free accessibility, MEDLINE is the most important bibliographic database in the field of biomedicine. It contains bibliographic citations and author abstracts from over 4,600 biomedical journals. Each citation is associated with a set of MeSH terms that describe the content of the item (Figure 1). Presently the database comprises over 12 million records dating back to 1966 [6]. MEDLINE is available for free searching on many websites of government and health agencies. One of the most popular is the NLM Web based product PubMed. There are also about 80 commercial products that provide access to MEDLINE.

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.

Title:
Improving the convenience of home-based interferon beta-1a therapy for multiple sclerosis.
Authors:
Lesaux J, Jadback G, Harraghy CE.
Abstract:
Subcutaneous interferon beta-1a (Rebif) therapy has been recognized as a significant advance in the treatment of relapsing-remitting multiple sclerosis (MS).
...
MeSH Terms:

- Adjuvants, Immunologic/therapeutic use*
- Adjuvants, Immunologic/administration & dosage
- Adult
- Home Nursing/methods*
- Human
- Injections, Subcutaneous
- Interferon-beta/therapeutic use*
- Interferon-beta/administration & dosage
- Middle Age
- Multiple Sclerosis, Relapsing-Remitting/nursing*
- Multiple Sclerosis, Relapsing-Remitting/drug therapy*
- Ontario
- Patient Compliance
- Patient Education
- Self Administration

….

Fig. 1. An example of a MEDLINE record. Only the title, authors, a part of the abstract and the MeSH terms (descriptors) fields are shown.

Medical Subject Headings (MeSH)

MeSH comprises NLM's controlled vocabulary and thesaurus used for indexing articles and for searching MeSH-indexed databases, including MEDLINE. It contains biomedical subject headings (descriptors), subheadings, and supplementary chemical terms. MeSH terms provide a consistent way to retrieve information that may use different terminology for the same concepts. MeSH organizes its descriptors in a hierarchical structure that permits searching at various levels of specificity from narrower to broader. This structure also provides an effective way for searchers to browse MeSH in order to find appropriate descriptors. A retrieval query is formed using MeSH terms to find items on a desired topic. Similarly, indexers normally use the most specific descriptors available to describe the subject content of an article. Problems with MeSH indexing may arise when not all important terminology in a

field is covered: when using descriptors, new concepts may be worded in a way that nonexpert users cannot readily identify. Problems may also arise because of inconsistency of human indexing [7]. The current MeSH includes more than 21,000 descriptors, over 132,000 Supplementary Concept Records (formerly Supplementary Chemical Records), and over 300,000 synonyms and related terms [8].

In our system, MeSH represents the set of biomedical concepts we are dealing with. In the first phase, we extract known relations between these concepts and in the second, we try to discover new relations between them.

Unified Medical Language System (UMLS)

Providing improved access to search terms as well as databases are goals of the Unified Medical Language System (UMLS) project that NLM began in 1986. The UMLS project was undertaken in order to provide a mechanism for linking diverse medical vocabularies as well as sources of information because the proliferation of disparate vocabularies, none of which was compatible with any other, was recognized as a significant impediment to the development of integrated applications. The project develops "Knowledge Sources" that can be used by a wide variety of applications programs to overcome retrieval problems caused by differences in terminology and the scattering of relevant information across many databases. There are now three UMLS Knowledge Sources: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon [9,10].

The Metathesaurus provides a uniform, integrated distribution format for terms from about 60 biomedical vocabularies and classifications used in patient records, administrative health data, bibliographic and full-text databases and expert systems.

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 and the permissible relationships among these types (e.g., "Virus" causes "Disease or Syndrome"). The Semantic Network, through its 134 semantic types, provides a consistent categorization of all concepts represented in the UMLS Metathesaurus.

The Lexicon contains syntactic information for many terms, component words, and English words, including verbs, which do not necessarily appear in the Metathesaurus.

UMLS research has made progress on some of the many research issues associated with interpretation of user queries, mapping between the language of different information sources, and medical indexing and retrieval techniques. Much of the serious investigation and prototype system development involving links between automated patient data and knowledge-based information has been performed using UMLS components [10].

In our system, we use UMLS information regarding the semantic types of the biomedical concepts and their co-occurrence in MEDLINE records.

Association rules

Association rules [11] were originally developed with the purpose of market-basket analysis, where it is of interest to find patterns of the form X→Y, with the intuitive meaning "baskets that contain X tend to contain Y". A basket corresponds to a single visit of a customer to a store and is called a transaction, while individual products in the basket are called items. The approach is general enough to apply to bibliographic databases, where transactions are documents and items are words or descriptors used for indexing the documents. Association rules here have the form Word1→Word2 or Descriptor1→Descriptor2 (e.g. Disease X (Multiple sclerosis)→Symptom Y (Optic neuritis)). Another example: Disease X (Multiple sclerosis)→Treatment Y (Interferon-beta).

4. System Description

Goal and Basic Premises

The system we developed is an interactive discovery support system for the field of medicine and is intended to be used as a generator of new, potentially meaningful relations between a starting, known concept of interest and other concepts.

The MEDLINE database is used heavily by biomedical researchers. Traditionally it has been used to check what is new in the literature on a particular topic of interest or to check if a medical discovery has already been published. In the latter case, the researcher has already made the discovery or at least has a general discovery idea, and just wants to check if someone else has already published that discovery. In addition, the various information retrieval systems used for searching MEDLINE are geared towards the task of searching for documents about a topic well known in advance. In contrast to the traditional use of MEDLINE, where it is used to check if a discovery is new or not, our system actively helps in the discovery process by generating potentially new discoveries and research ideas by analyzing the MEDLINE database.

When building our discovery support system, we started with the same basic ideas as Swanson. However, the methods we use for discovering new relations between concepts are different as well as the software implementation of these methods.

We use the major MeSH descriptors assigned to a MEDLINE record as a representation of the contents of the article. Some of the MeSH descriptors are designated as major (followed by an asterisk in the MEDLINE record). Major descriptors are those that form the main topic of the article. See Figure 1 for an example of a MEDLINE record and MeSH descriptors.

We use association rules [11] between pairs of medical concepts as a method to determine which concepts are related to a given starting concept. In our system an association rule of the form

X→Y (confidence, support)

means that in *confidence* percent of articles containing X, Y is also present and that there are *support* number of articles containing both X and Y. In other words, we take

concept co-occurrence as an indication of a relation between concepts. 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. The system does not try to find out the kind of relation currently. This cannot be done directly by using the MeSH descriptors assigned to an article because there is no explicit information about the relation between the descriptors stored in the MEDLINE record. However, the system prepares a query on demand and retrieves the MEDLINE records containing both X and Y. By reading the titles and abstracts of these records, the user can determine the nature of the relation.

Discovery Algorithm

We calculated all the associations between the major MeSH descriptors for two subsets of MEDLINE citations. We did this regardless of the confidence and support values and for two MEDLINE time segments: 1990-1995 and 1996-1999. The calculated associations are stored in a database management system: there are currently more than 11.000.000 associations in the rule base. The calculation of the association rules was much simplified by the use of the data contained in the UMLS, especially the co-occurrence files. Actually, for these calculations it was not necessary to access the full MEDLINE records at all.

The large association rule base is a foundation upon which the algorithm for discovering new relations between concepts proceeds as described in Table 1. The main idea is to first find all the concepts Y related to the starting concept X (e.g. if X is a disease then Y can be pathological functions, symptoms, ...). Then all the concepts Z related to Y are found (e.g. if Y is a pathological function, Z can be a chemical regulating that function). As a last step, 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 user of the system should then evaluate the proposed (X, Z) pairs and select among them those that deserve 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.

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→Y.
3. Find all concepts Z such that there is an association rule Y→Z.
4. Eliminate those Z for which an association X→Z already exists.
5. The remaining Z concepts are candidates for a new relation between X and Z.

Because in MEDLINE each X concept can be associated with many Y concepts, each of which can be associated to 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 default filtering that cannot be relaxed is that only the associations between major MeSH headings are considered by the system. The other filtering possibilities are optional and can be interactively enforced by the user of the system.

The related concepts can be limited by the semantic type to which they belong. Each MeSH descriptor belongs to one or more semantic types. For example, if the starting concept X is a disease (semantic type *disease or syndrome*) then the user can request that Y concepts are of semantic type *pathologic function* and that Z concepts are of semantic type *pharmacologic substance*. Consequently, the system will only consider chains of associations of the form: *disease or syndrome→pathologic function→pharmacologic substance.* 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. In fact, all of the filtering options can be interactively set alone or several of them can be selected in combination.

Because the usefulness of the system relies to a large degree on human judgment, special attention is paid to the order in which the candidate related concepts are presented to the user of the system. Thus, 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 decreasing association rule confidence, but it is also possible to order by support or semantic type.

User Interface

During a discovery session, the user has to browse and evaluate many potential discovery candidates. Therefore, user-friendliness and short response time were our most important goals when designing the user interface of the discovery support system (Figure 2). Additionally, the user needs a standard web browser to access MEDLINE through the Entrez system, through which it is also possible to search the GenBank, SwissProt, OMIM and other databases. We will now describe the elements of the user interface, its use in the discovery process and its integration with the Entrez system.

Searching for a Starting Concept X

The user initiates a discovery session by searching for a starting concept X, which is usually from his own research area. The query is performed with a *query-by-form* method using the same screen form for browsing data and for searching. The user can specify a full or partial name in the *Str* field. When specifying a partial name, the *%* (percent) sign is a wild-card character and is a replacement for any character string, including an empty one. After entering the search criteria, the user presses the *Query* button to retrieve the results, which are shown in the *Starting Concept* frame and the semantic types it belongs to are shown in the frame to the right.

When specifying a partial name, it is possible that more then one concept matches the search criteria. In that case, the system shows the first matching concept. However, by using the navigation buttons (<<, <, >,>>), it is possible to show the other concepts as well.

The fields above the concept name show the frequency of occurrence of the concept in MEDLINE when used as a major MeSH descriptor for two time intervals (1990-1995) and (1996-1999).

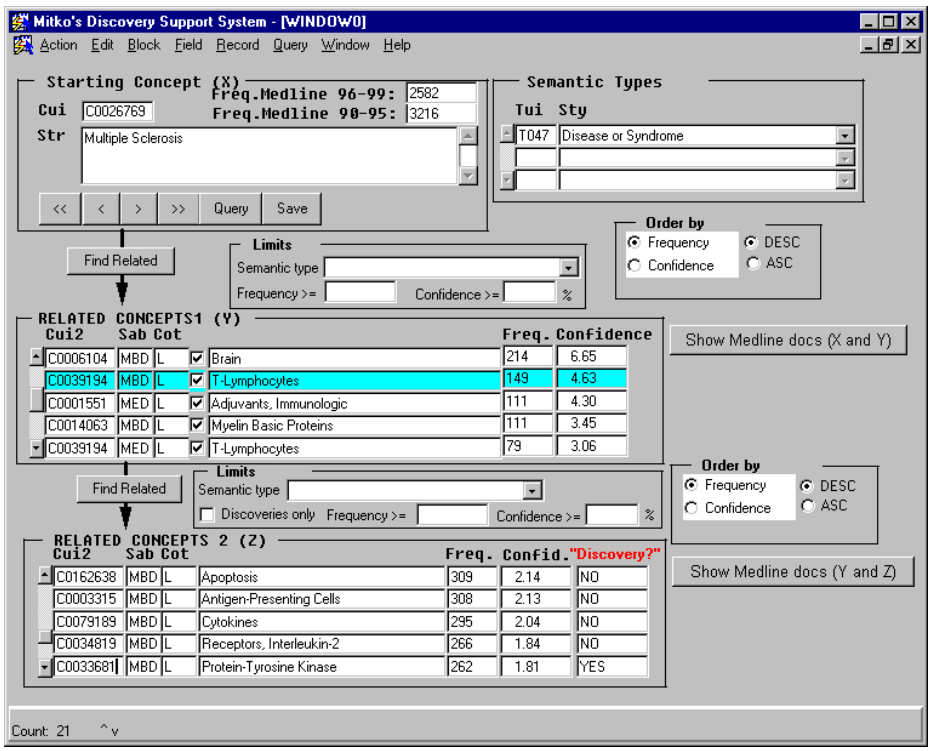

Fig. 2. The user interface of the interactive discovery support system.

Finding the Related Concepts Y

The concepts Y related to the starting concept are found by pressing the *Find Related* button that is under the *Starting Concept* frame. Before finding related concepts, the user can specify limits such as the semantic type of the related concepts or the minimal confidence and frequency (support) of the association rules. This is done in the upper *Limits* frame. In addition, the order in which the related concepts are presented can be specified in the upper *Order by* frame. It can be descending or ascending by confidence or frequency (support). When the user wants to try a new limiting and ordering combination, he/she has to press the *Find Related* button again.

The related concepts are presented in the *Related Concepts1* frame. The *Sab* (Source abbreviation) column shows the source of relationship (MBD - MEDLINE 1990-1995, and MED - MEDLINE 1996-1999). Apart from the related concept names, the frequency of co-occurrence in the particular segment is shown as well as the confidence of the association rule between the starting concept and the related concept. The user can browse through the list of related concepts and select those that

need further investigation. This is done by selecting the check box to the left of the related concept name.

Finding the Related Concepts Z

Now the user can press the *Find Related* button, which is under the *RELATED CONCEPTS1* frame. This action will find the Z concepts related to the Y concepts found in the previous step and show them in the *RELATED CONCEPTS2* frame. As described earlier, the user can specify limits and the order of the related concepts.

The frame *RELATED CONCEPTS2* contains an important additional field designated as *"Discovery?"*. The value of this field is YES if a relation (association) between the starting concept X and the current concept Z does not exist in the appropriate MEDLINE segment and *NO* if such a relation exists. In other words, this field shows if the relation between the starting concept X and the current concept Z is considered a potential discovery by the system. Because of the ten year MEDLINE interval used, it is possible that some of the potential discoveries have been described at an earlier time point, and thus are not good candidates for a discovery. In any case, the judgment of a human expert (hopefully the user of the system) is needed to verify how plausible the potential discoveries offered by the system are. The user can browse through the list of these potential discoveries or she can select different set of Y concepts and try again to find some potential discoveries. It is possible to limit the Z concepts to only those considered potential discoveries by checking the check box *Discoveries only* in the lower *Limits* frame.

When one starting concept has been dealt with, another one can be searched for and the whole procedure can be repeated. The user can interactively guide the discovery process by selecting promising concepts and by setting various limits.

Searching and Browsing related MEDLINE records

To make the evaluation of the proposed potential discoveries easier for the user, the system provides the ability to search and display the MEDLINE records related to the concept currently under investigation. The user can read these records, decide which relations deserve further attention and guide the discovery process accordingly.

The search and display of MEDLINE records is accomplished by pressing either the *Show MEDLINE docs (X and Y)* or the *Show MEDLINE docs (Y and Z)* buttons. The first one searches for MEDLINE records containing both the starting concept X and the current concept Y. Similarly, the second button searches for records containing both the current Y and Z. When the user presses one of these buttons, the discovery system does the following: 1) prepares an appropriate search request, 2) starts a standard web browser, 3) connects to the search system Entrez, and 4) runs the prepared search request. The user can then browse through the resulting records and change the search request if necessary. With some basic knowledge of the Entrez system, it is also possible to display the related proteins and nucleotide sequences.

Implementation

The association rule base and the other necessary tables needed in the discovery process are stored in an Oracle relational database management system on a UNIX server. We have developed three versions of the end user program. The first one was developed using Oracle Forms6i. It communicates with the database server in a client/server manner over a TCP/IP network, which can also be the Internet. We use this version only in-house because it requires software installation on the user's computer. To make the system widely available, we developed the second and third versions. They are both Web based and require only a Web browser on the user side. The second version, which has functionality identical to the first, is a three-tier application with the user tier being implemented in Java. The third version is a CGI-BIN style application and does not require Java support on the user's side.

5. Evaluation and Results

We evaluated the system in three different ways: first, by checking the medical meaning of the relations extracted by association rule mining; second, by a statistical evaluation in which we checked how effective the system was in predicting new relations; and third, by using the system to discover the candidate gene for a genetic disease.

Medical Meaning of Related Concepts

This evaluation was conducted by a medical doctor (Borut Peterlin, one of the co-authors of this chapter). He used the system to check if the Y concepts, which the system had found to be related to the starting concept X, were medically correct. In other words, we wanted to determine whether the association rules are successful in extracting known relations between biomedical concepts from the MEDLINE database. This is very important in our approach because by combining known relations, our system proposes potentially new relations. If the system failed to extract the known relations, then we could not expect it to discover new relations either.

The doctor selected as a starting concept *multiple sclerosis (MS)*, which can be defined as a demyelinating disease of the central nervous system of putative autoimmune origin. Then he used the system to find the related concepts Y. Below is a list of the first 20 concepts related (associated) with MS ordered by decreasing support, with their type and a short description of the nature of the association:

1. *MRI magnetic resonance imaging (diagnostics).* MRI is nowadays the method of choice to confirm the diagnosis of MS. It has a relatively high sensitivity (80%) and is noninvasive.
2. *Brain (anatomical structure – organ involved).* It simply reflects the anatomical structure of the central nervous system often affected in MS.
3. *Interferon (treatment).* A prophylactic drug nowadays used to reduce the rate of attacks – by 30%.

4. *T-lymphocytes(pathogenesis).* T lymphocytes regulate humoral immune responses and are found in abundance within MS lesions. It is believed that in MS, a T-cell mediated, autoimmune inflammatory reaction, at least as a mechanism for sustaining the inflammation, is involved.
5. *Myelin basic protein (MBP)(pathogenesis).* This structural component of myelin is potentially involved in the pathogenesis of MS. Antibodies to MBP have been found in both the serum and cerebrospinal fluid (CSF) of MS patients, and these antibodies, along with T cells that are reactive to MBP, increase with disease activity.
6. *Optic neuritis (symptoms).* Optic neuritis is one of the most common symptoms of MS (in 40% of patients).
7. *Autoimmune diseases (disease categories).* Due to involvement of immunocompetent cells, association with certain HLA types, oligoclonal bands in liquor, abnormal subsets of T-cells, animal model of MS – EAE (experimental autoimmune encephalomyelitis) is an immune mediated disease, and autoimmunity is considered to be an important etiological factor in MS.

8.-20. *Immunosupresives (treatment), IgG (diagnostics), encephalomyelitis (symptoms), cognition disorders (symptoms), VEP (diagnostics), citokines (pathogenesis), TNF (pathogenesis), spinal cord (anatomical structure – organ involved), methylprednison (treatment), receptors-Tcells (pathogenesis), myelin protein (pathogenesis), psychological adjustment (treatment), demyelinating disorders (disease categories).*

Among the 20 concepts analyzed, 6 are related to pathogenesis, 4 to treatment, 3 to diagnostic methods, 3 to symptoms, 2 to target organs-anatomical structures and 2 are related to general disease categories.

We conclude that the concepts found as related by the system are associated with the current main focus of medical endeavors in the fie ld of MS, which is still oriented to treatment and therefore to better understanding of pathogenesis.

*Table 2 - The results of the prediction of new relationships between medical concepts in the newer MEDLINE segment (1996-1999) based on the older segment (1990-1995) using the system. The column names ending with **1** are for the AVGS constraint and those ending with **2** for the 2*AVGS constraint. The columns have the following meaning: **n** - all the relationships that can be predicted; **k** - new relationships in the newer segment that were not present in the older segment; **m** - predicted relationships based on the older segment; **l** - successfully predicted relationships; **p** - probability of achieving **l** or more successfully predicted relations by chance; **r** - the number of successfully predicted relations by chance alone.*

Disease	***n***	***k***	***m1***	***l1***	***p1***	***r1***	***m2***	***l2***	***p2***	***r2***
Multiple Sclerosis (MS)	15965	635	6848	521	0	272	3151	366	0	125
Temporal Arteritis (TA)	17190	187	4735	148	0	52	1157	72	0	16
Melanoma (ML)	15336	692	6272	560	0	283	2812	392	0	127
Parkinson Disease (PD)	15966	594	5995	477	0	223	2322	309	0	86
Incontinentia Pigmenti (IP)	17504	44	3435	37	0	9	873	23	0	2
Chondrodysplasia Punctata (CP)	17422	18	2864	15	0	3	1046	9	0.00000016	1
Charcot-Marie-Tooth Disease (CMT)	17355	131	3150	105	0	24	1019	66	0	8
Focal Dermal Hypoplasia (FDH)	17527	23	1511	14	0	2	610	8	0.00000037	1
Noonan Syndrome (NS)	17384	68	3015	59	0	12	536	23	0	2
Ectodermal Dysplasia (ED)	17322	124	3301	96	0	24	967	45	0	7

Statistical Evaluation

The goal of the statistical evaluation was to see how many of the potential discoveries predicted by the system at some point in time become realized at a later time. For us, a potential discovery is a relationship between two concepts proposed by our system, but not present in MEDLINE at some point in time. We consider the potential discovery realized if the two concepts later appear together in a document in the MEDLINE database. In other words, the goal of the evaluation was to see how good our system was in predicting what discoveries would be made in the future.

We approached this goal by first dividing the MEDLINE database and the corresponding association rules into two segments according to the publication date of the documents stored: the older segment is from 1990 to 1995 and the newer segment is from 1996 to 1999. We then analyzed ten diseases, which are listed in Table 2.

Here we will give a discussion of the analysis of *Multiple sclerosis (MS).* MS appears in 2582 documents in the older segment designated as a major MeSH descriptor. It is related to 1610 distinct concepts. When analyzing the old segment, the system proposed 15617 concepts as potential discoveries. MS is related to 635 new concepts in the new segment that it was not related to in the old segment. Our system successfully predicted 99.5% (632 out of 635) realized discoveries in the new segment. However, only 4% (632 out of 15617) of the proposed potential discoveries got realized. It should be stressed that MS was not related to 15965 out of 17575 distinct concepts appearing in the older segment. The system proposed 97.8% of the concepts MS was not yet related to as potential discoveries. The conclusion is that without using limits on the strength of relationship, the system is very successful at predicting future discoveries, but proposes far too many potential discoveries.

We then repeated the evaluation with two values for thresholds on the support level of the association rules. In one case the threshold was set to the average support of the associations between one concept and the others (AVGS) and in the other case it was set to 2*AVGS. Only associations with support greater than or equal to the threshold were taken into account. The number of proposed potential discoveries dropped from 15617 without thresholds to 6848 for AVGS and to 3151 for the 2*AVGS threshold. The percent of successfully predicted and realized discoveries dropped from 99.5% (632 of 635) without thresholds to 82.0% (521 of 635) for AVGS and to 57.6% (366 of 635) for 2*AVGS. However, the ratio of realized to proposed potential discoveries improved from 4% (632 out of 15617) without thresholds to 7.6% (521 of 6848) for AVGS and to 11.6% (366 of 3151) for 2*AVGS.

The results of the statistical evaluation for the ten selected diseases are in Table 2. The values obtained by our system were tested against the null hypotheses of random hits. Or to put it another way, we wanted to check whether the number of correct predictions obtained by our system could have occurred by chance alone. This is done in the following way.

Let n be the number of all possible relationships that the system could predict based on the older MEDLINE segment. Of these, k actually appear in the new segment (successful predictions), and n-k do not (nonsuccessful predictions). Let m be

the number of actual predictions made by the system, of which l are successful, and m-l nonsuccessful. The probability of such an event is

$$\frac{\binom{k}{l}\binom{n-k}{m-l}}{\binom{n}{m}}.$$

This distribution is known as the hypergeometric distribution. The question now is whether l successful predictions represent a statistically significant result. In other words, we need the probability of obtaining l or more successful predictions if we were predicting completely at random. This probability is given by

$$p = \sum_{i=l}^{k} \frac{\binom{k}{i}\binom{n-k}{m-i}}{\binom{n}{m}},$$

or, equivalently

$$p = 1 - \sum_{i=0}^{l-1} \frac{\binom{k}{i}\binom{n-k}{m-i}}{\binom{n}{m}}.$$

Table 2 shows p values for given n, m, k and l for AVGS and 2*AVGS constraints respectively. Zeros in the p value columns actually mean that the probability is less than 10^{-16}. If the predictions were random, we would expect $\frac{m}{n}k$ of them to be successful.

Table 3 shows a summary of the relationship prediction results in terms of precision and recall for the AVGS and 2*AVGS constraints respectively. In our context, we define precision and recall as follows. Precision is the percentage of correctly predicted new relationships among all predicted relationships (***l/m*** from Table 2). Recall is the percentage of correctly predicted new relationships among all new relationships (***l/k*** from Table 2). For the AVGS constraint, the precision ranges from 0.5% to 8.9% with an average of 3.8%, and the recall ranges from 60.9% to 86.8% with an average of 79.5%. For the 2*AVGS constraint, the average precision increases to 6.5% and the average recall falls to 46.2%. However, the increase in precision is equal to the decrease in recall and is around 71%. The column ***Better then***

random in Table 3 shows how much better are the predictions of the system compared with random predictions (***l/r*** from Table 2). The value of this column is 3.8 for the AVGS constraint and 6.9 for 2*AVGS.

*Table 3 – Summary relationship prediction results for the AVGS and 2*AVGS constraints respectively:* ***Precision*** *(correctly predicted among all predicted),* ***Recall*** *(correctly predicted among all relationships),* ***Better then random*** *(correct predictions of the system divided by random correct predictions).*

Disease	***AVGS***			***2*AVGS***		
	Precision	***Recall***	***Better then random***	***Precision***	***Recall***	***Better then random***
MS	7.6%	82.0%	1.9	11.6%	57.6%	2.9
TA	3.1%	79.1%	2.8	6.2%	38.5%	4.5
ML	8.9%	80.9%	2.0	13.9%	56.6%	3.1
PD	8.0%	80.3%	2.1	13.3%	52.0%	3.6
IP	1.1%	84.1%	4.1	2.6%	52.3%	11.5
CP	0.5%	83.3%	5.0	0.9%	50.0%	9.0
CMT	3.3%	80.2%	4.4	6.5%	50.4%	8.3
FDH	0.9%	60.9%	7.0	1.3%	34.8%	8.0
NS	2.0%	86.8%	4.9	4.3%	33.8%	11.5
ED	2.9%	77.4%	4.0	4.7%	36.3%	6.4
Average:	**3.8%**	**79.5%**	**3.8**	**6.5%**	**46.2%**	**6.9**

Disease Candidate Gene (re)Identification

In the preceding section, we showed that our system predicts new relations between medical concepts with statistical significance better than chance. This time we wanted to evaluate the system for the task of candidate gene discovery for a disease. For this purpose, we selected the incontinentia pigmenti disease.

Familial incontinentia pigmenti (IP; MIM 308310) is a genodermatosis that segregates as an X-linked dominant disorder and is usually lethal prenatally in males. In affected females, it causes highly variable abnormalities of the skin, hair, nails, teeth, eyes and central nervous system. The gene for IP was mapped in the terminal part of the long arm of the X chromosome (Xq28) in 1994 [12]. The pathogenesis of the disease is not known yet; however, the immune and haematopoetic systems seem to play an important role [13, 14, 15].

In 2000, the gene NEMO (NF-kappaB essential modulator/IKKγ) mutated in patients with IP was identified via positional cloning approach [16]. NEMO is required for activation of the NF-kappaB transcription factor, which is involved in numerous immune, inflammatory and apoptotic mechanisms.

The NEMO gene was cloned in 1998 and localized in the Xq28 region of the human chromosome X in 1999 [17].

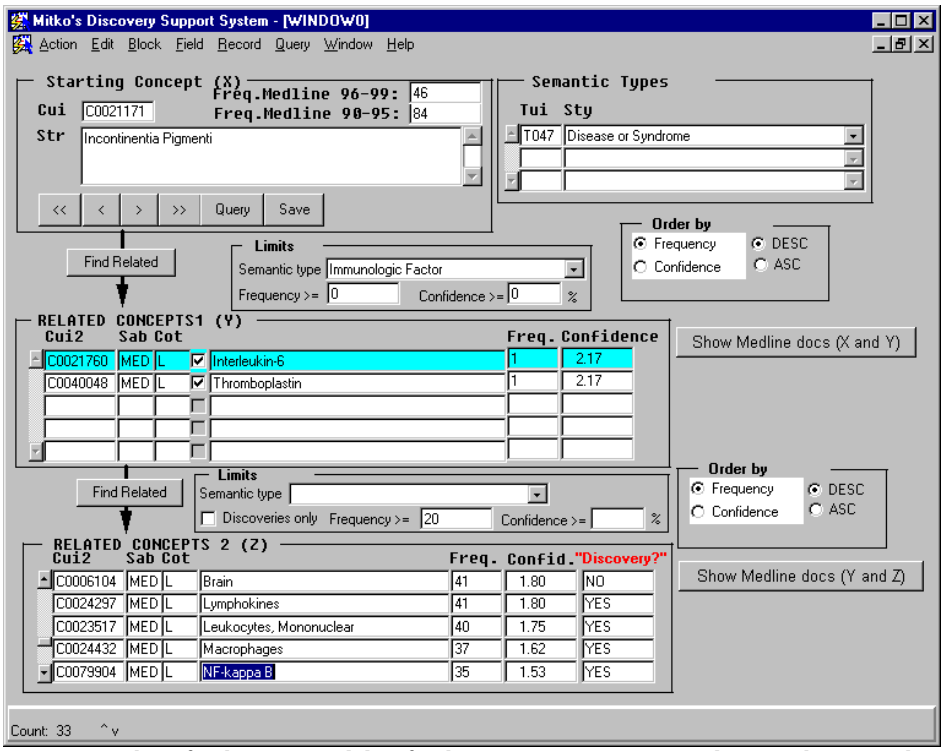

Fig. 3. Searching for the gene candidate for the *Incontinentia Pigmenti* disease. The system has discovered a potentially new relationship between *Incontinentia Pigmenti* and the *NF-kappa B* transcription factor through the intermediate concept *Interleukin-6*.

More precisely, our aim regarding IP was to find out whether NEMO could have been identified by our discovery support system using the data that was available before NEMO was officially identified as the gene for IP.

As the starting concept (X) the name of the disease has been entered – Incontinentia pigmenti (Figure 3). Reflecting the fact that the immune system seems to play an important role in the pathogenesis of IP, related concepts of the first order (Y) were limited by the semantic type *immune factor*. Only two related concepts were found: *Interleukin-6* (IL-6) and *Thromboplastin*. IL-6 is a cytokine and has an important role in the development of the inflammatory response, the differentiation and activation of cells of the hematopoietic lineage, and the regulation of nerve cell and bone cell functions. By pressing the *Show MEDLINE docs (X and Y)*, the document containing both IP and IL-6 is retrieved and displayed [14]. It reports that increased serum levels of IL-6 have been found in an IP patient, who in addition to IP also demonstrated symptoms of Behcet disease.

In the next step, we searched for concepts related to IL-6 (concepts Z). The column *"Discovery?"* shows if there might be a potentially new relation between the starting concept X and the current concept Z. The value in this field is *YES* if there are no MEDLINE documents containing both the X and Z concepts (as major MeSH descriptors) in the corresponding MEDLINE segment. Among the hits with higher co-

occurrence frequency, we identified three hits pathogenetically potentially related to IP: NF-kappa B, apoptosis and vascular endothelium.

NF-kappa B is a transcription factor involved in metabolic pathways related to immune system, inflammation and apoptosis. The MEDLINE documents containing both IL-6 and NF-kappa B show that NF-kappa B is engaged in the activation and is the central mediator of the expression of IL-6 (e.g. [18]). Therefore, NF-kappa B is very interesting regarding IP.

In the next step, we used the fact that the gene for IP is located at the Xq28 chromosomal region, which has been known since 1994 [12]. When the corresponding button labeled *Show MEDLINE Docs* is pressed, an external Web browser is started and a search request is executed on the PubMed database to display the concepts that are currently under consideration. Figure 4. shows the results of the search request for documents containing the transcription factor *NF-kappa B* and chromosome location *Xq28*. The first article [16] reports the discovery of the NEMO gene as responsible for IP. However, this article was published in 2000. The second article [17] is very important in our procedure because it states that the NEMO gene, which is necessary for the activation of the transcription factor NF-kappa B, is localized in the Xq28 region. This article does not make any reference to IP and was published in 1999. From this we can conclude that by using our system together with some MEDLINE searching, it would have been possible to predict the NEMO gene as responsible for IP in 1999. In this example, the path from the IP disease to the NEMO gene went through two intermediate concepts: IL-6 and NF-kappa B.

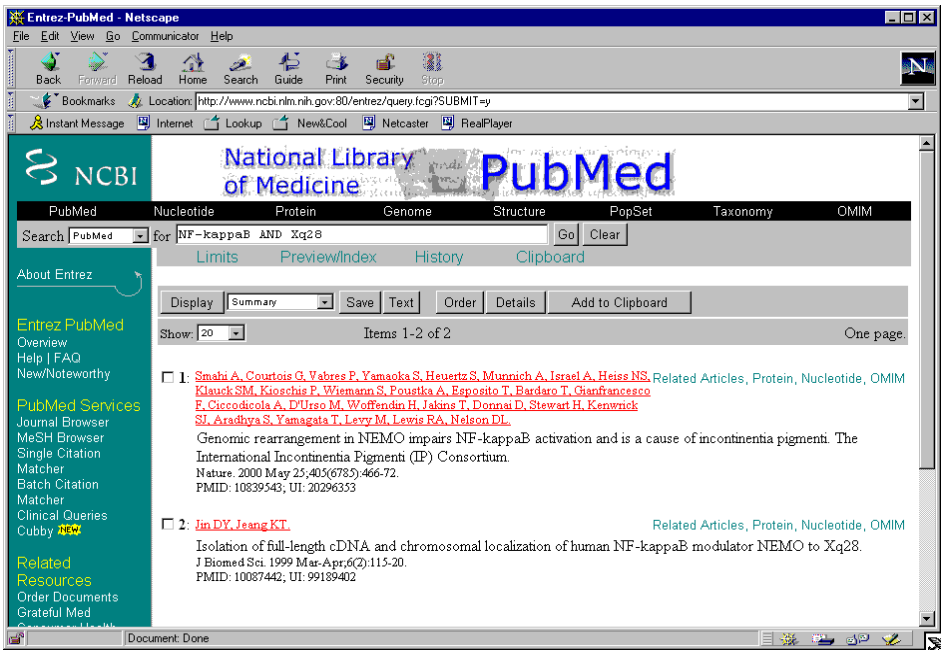

Fig. 4. The MEDLINE articles containing both the transcription factor *NF-kappa B* and the chromosomal region *Xq28*.

6. Discussion and Further Work

We have presented an interactive discovery support system (available at http://www.mf.uni-lj.si/bitola) for the field of medicine. For a given starting medical concept, it discovers new, potentially meaningful relations with other concepts that have not been published in the medical literature before. The proposed relations then need to be evaluated and verified by a qualified medical professional.

As a measure of the relation between concepts, we use association rules calculated from the MEDLINE bibliographic database. We were in a dilemma whether to use the X→Y or Y→X direction of the association rule when finding concepts related to X. Because we have only binary associations, the direction comes into play only when limiting and ordering the related concepts. We selected the X→Y direction based on the intuition that the Y concepts that appear most often in documents regarding X have the strongest relation to X. However, there are cases where some Y concepts might appear infrequently in X documents, but X might appear in almost all Y documents. In this case, there is a strong association in the Y→X direction that should also be considered. We plan to further investigate this issue. One possibility might be to develop a heuristic approach in which sometimes an X→Y association is used and sometimes a Y→X one. Another possibility is to use some kind of a composite measure that takes into account both directions of the associations.

We have not used MEDLINE directly, but rather we use the UMLS. It simplifies the calculations considerably; however, it introduces considerable limitations into our system as well: we can calculate only binary association rules; the association rules are only between major MeSH headings; and we are limited to only two time intervals. Currently in MeSH, only a few dozen genes and other sequences are present. We plan to calculate direct association rules between the MeSH headings and a considerable number of molecular biology sequences and thereby increasing the functionality of the system significantly. To alleviate the above mentioned weaknesses, we are currently developing a new version of the system with these features: it analyzes full MEDLINE, expands the set of concepts with all confirmed human genes, and includes additional domain knowledge, such as the chromosomal location of the diseases and genes.

As part of the evaluation of the system, a medical doctor confirmed that most of the relations between concepts found by our association rules are meaningful. We demonstrated a successful application of the system for predicting a gene candidate for the incontinentia pigmenti disease, by using the information known prior to the discovery of the gene. In the statistical evaluation, the system proved to be successful at predicting future discoveries. However, this came at the expense of generating a large number of potential discoveries that have to be judged and verified by the user of the system. The statistical evaluation also showed that properly set thresholds are crucial for successful use of the system. Thus, we plan to work on setting good default values for the thresholds that can be changed by the user if necessary.

We believe that literature based discovery support systems, such as ours, will help researchers make some important biomedical discoveries in the future.

7. Acknowledgments

The authors would like to thank Tom Rindflesch and Alan Aronson for providing valuable comments and insights.

References

1. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986 Autumn;30(1):7-18.
2. Swanson, D.R.: Migraine and magnesium: eleven neglected connections. Perspect Biol Med. 1988 Summer;31(4):526-57.
3. Swanson, D.R., Smalheiser, N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. 91 (1997) 183-203.
4. Gordon MD, Lindsay RK. 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 1996; 47(2):116-128.
5. Weeber M, Klein H, Aronson AR, Mork JG, Jong-Van Den Berg L, Vos R. Text-based discovery in biomedicine: the architecture of the DAD-system. Proc AMIA Symp. 2000;(20 Suppl):903-7.
6. U.S. National Library of Medicine - MEDLINE. http://www.nlm.nih.gov/databases/databases_MEDLINE.html<28.01.2003>
7. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc. 1983 Apr;71(2):176-83.
8. Medical Subject Headings - MeSH. http://www.nlm.nih.gov/mesh/<28.01.2003>
9. Unified Medical Language System – UMLS. http://www.nlm.nih.gov/research/umls/<28.01.2003>
10. Humphreys, B.L., Lindberg, D.A.B., Schoolman, H.M., Barnett, G.O.: The Unified Medical Language System: an informatics research collaboration. JAMIA 1998;5(1):1-11.
11. 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)
12. Smahi A, Hyden-Granskog C, Peterlin B et al. The gene for the familial form of incontinentia pigmenti (IP2) maps to the distal part of Xq28. Hum Mol Genet 1994; 3:273-8.
13. Roberts JL, Morrow B, Vega-Rich C, Salafia CM, Nitowsky HM. Incontinentia pigmenti in a newborn male infant with DNA confirmation.Am J Med Genet 1998;75:159-63.
14. Endoh M, Yokozeki H, Maruyama R, Matsunaga T, Katayama I, Nishioka K. Incontinentia pigmenti and Behcet's disease: a case of impaired neutrophil chemotaxis. Dermatology 1996;192:285-7.

15. Dahl MV, Matula G, Leonards R, Tuffanelli DL. Incontinentia pigmenti and defective neutrophil chemotaxis. Arch Dermatol 1975;111:1603-5.
16. The International Incontinentia Pigmenti Consortium. Genomic rearrangement in NEMO impairs NF-?B activation and is a cause of incontinentia pigmenti. Nature 2000;405:466-72.
17. Jin DY, Jeang KT. Isolation of full-length cDNA and chromosomal localization of human NF-kappaB modulator NEMO to Xq28. J Biomed Sci. 1999 Mar-Apr;6(2):115-20.
18. Vanden Berghe W, De Bosscher K, Boone E, Plaisance S, Haegeman G. The nuclear factor-kappaB engages CBP/p300 and histone acetyltransferase activity for transcriptional activation of the interleukin-6 gene promoter. J Biol Chem. 1999 Nov 5;274(45):32091-8.