LINKING BIOMEDICAL LANGUAGE INFORMATION AND KNOWLEDGE RESOURCES: GO AND UMLS

I.N. SARKAR[*,1], M.N. CANTOR[*,1], R. GELMAN[1], F. HARTEL[2] AND Y.A. LUSSIER[§,1,3]

1- Department of Medical Informatics, Columbia University College of Physicians and Surgeons, New York, NY 10032 USA

2- Center for Bioinformatics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892 USA

3- Department of Medicine, Columbia University College of Physicians and Surgeons, New York, NY 10032 USA

Integration of various informatics terminologies will be an essential activity towards supporting the advancement of both the biomedical and clinical sciences. The GO consortium has developed an impressive collection of biomedical terms specific to genes and proteins in a variety of organisms. The UMLS is a composite collection of various medical terminologies, pioneered by the National Library of Medicine. In the present study, we examine a variety of techniques for mapping terms from one terminology (GO) to another (UMLS), and describe their respective performances for a small, curated data set attained from the National Cancer Institute, which had precision values ranging from 30% (100% recall) to 95% (74% recall). Based on each technique's performance, we comment on how each can be used to enrich an existing terminology (UMLS) in future studies and how linking biological terminologies to UMLS differs from linking medical terminologies.

1 Introduction

1.1 Need to Link Structured Information Resources

As is often the case in an advancing science, there are divisions between various people working towards a common goal. This is no different in the information sciences – whether in medical or bio- informatics[1,2,3]. The quick advancement of technology, both computational and laboratory, has enabled the recent integration of medical and bio- informatics[1,2]. To this end, we have explored how various methods can be used to help integrate the essential terminologies of these diverse, yet, similar fields.

The respective terminologies that serve the medical and biological sciences communities are, by themselves, of great importance to each individual field. Links between the two fields, however, are growing, as medicine increasingly incorporates basic biological science advances into clinical practice, and biologists or bioinformaticians validate their experiments using real patient data. These growing interactions necessitate a standardized method for communicating results between the fields.

---

* These authors contributed equally to the work

§ Corresponding author

Among the medical terminologies, the Unified Medical Language System (UMLS)[4] sets the standard for breadth of inclusion. The included terminologies, however, are generally focused on clinical medicine, so representation of more basic biological terms is often lacking. There is a need for representation of gene and gene products, such as those from the Human Genome Project (HGP), in the UMLS. There has been initial exploration into strategies to map genomic knowledge into the UMLS[5]. Aside from mapping specific terms from projects such as the HGP, however, there has been minimal documented speculation about the inclusion of entire biological terminologies, such as the Gene Ontology (GO)[6, 7].

Before terminologies such as GO are included within the UMLS, there needs to be an exploration into methodologies that may be utilized for the augmentation process. Analyses of the use of various text mining and information extraction techniques need to be performed. In the present study, we provide a preliminary analysis using a variety of applicable methodologies.

1.2 Not Really a New Art, but Still the Same Problems

Mapping of various medical terminologies to the UMLS has been studied extensively.[7,8,9,10,11,12] The UMLS 'Lexical Tools', provided by the National Library of Medicine (NLM)[13], offer some assistance with mapping of terms, but with varying results. However, the attempted methods have demonstrated limited success, generally able to map 13 - 60% of terms.[9,10]

Part of the complexity lies in the variety of ways that a single concept may be represented. As disparate systems often use the same information resources, it is imperative that redundancy be kept to a minimum. However, this lends itself to a great challenge when trying to augment composite terminologies, such as the UMLS, with highly structured terminologies such as GO.

Additionally, different terminologies may represent the same concept in very different ways. There may be future endeavors that will incorporate intelligent Natural Language Processing (NLP) techniques that will enable such disparities to be resolved[12]. Indeed, there has been some investigation of using the UMLS as a resource for language processing systems.[15,16]

However, using various non-NLP mapping techniques it should be possible to identify the correct concept classification of a significant portion of the non-UMLS terms. To date, there have been no published results of mapping GO terms to the UMLS. There have been some recent efforts in annotating genes from biomedical literature using the Gene Ontology.[17] Admittedly, GO was designed for the annotation of genes and gene products. For this reason, most of the recent focus has been on the annotation of gene and gene products in the biomedical literature and not in mapping terminologies. However, there has been some previous investigation into how other biomedical resources such as OMIM and GENBANK can be mapped

to the medical information structures.[18] This previous work demonstrates that there is a desire and a need for the integration of the various biomedical (clinical and basic science) terminologies.

2 Materials

2.1 The Unified Medical Language System

We used the 2001 version of the UMLS, created and maintained by the National Library of Medicine. The NLM began constructing the UMLS in 1986, to facilitate not only the retrieval and integration of information from diverse biomedical sources but also the linkage of disparate information systems.[4] The 2001 version of the UMLS consists of about 800,000 unique concepts (797,359) from over 60 diverse terminologies.[20] Each individual concept is represented in the UMLS by a 'Concept Unique Identifier' (CUI). There may be multiple text string variants (UMLS terms) affiliated with each CUI. The variants are identified by 'String Unique Identifiers' (SUIs). In the 2001 UMLS there are 1,728,075 SUIs.

The principal component of the UMLS used in the present study is the Metathesaurus. The UMLS Metathesaurus links "terminology and concepts from a range of vocabularies and classifications" through the UMLS Semantic Network, which provides a "consistent high level of categorization" to the individual concepts.[4] The UMLS, while generally focused on medical terminologies, does contain some gene and gene product names. These are mostly a result of the Medical Subject Headings initiative for indexing biomedical articles.[21,22]

2.2 Gene Ontology

The Gene Ontology consortium is focused on developing structured, coded, vocabularies for molecular function, biological processes, and cellular components that can be used across species.[6] Significantly, the vocabularies are independent of the associations between specific gene products and GO terms, leading to flexibility and precision in the use of the framework. The focus of the present study involved the May 2001 version of the GO. Each GO concept is represented by one unique text string also known as GO term. We refer to each GO concept by their identifying number (GOID).*2.3 Gold Standard*

The National Cancer Institute (NCI) provided us with several files containing mappings between a subset of GO (NCI-GO version May 2001) and the UMLS to their internal metathesaurus. From these files, we derived a subset of 332 distinct GOIDs that had been mapped to CUIs from the UMLS.

For the present study, we treated these distinct 332 GOID-CUI pairs as our gold standard (GS). We note that the GS contained 314 distinct CUIs, indicating that some CUIs mapped to more than one GOID. Since mapping methods make extensive use of the UMLS terms related to the CUIs, it is noteworthy to mention that 6,113 SUIs are associated to these 314 individual CUIs.

2.4 Applications and Scripts

All the applications and scripts to implement the methods discussed in this paper were written in C, C++, Java, and Perl. Additionally, Lexical Tools, the MMTx tool obtained from the National Library of Medicine, and applications associated with the National Center for Biotechnology Information's BLAST were also used. Most applications were run on a Sun machine running the SunOS 5.8 operating system. MMTx was run on a Personal Computer running Microsoft Windows 2000.

3 Methods

3.1 Approaches to Mapping Terminologies

The final goal of this project is to accurately capture and map the largest possible subset of the GOIDs to the UMLS CUIs. In order to find the most effective method for doing so, we evaluated four different coupling approaches to matching the two vocabularies. Because we did not use any machine learning methods a training set was not required. For all the methods implemented, we used the textual, flat file representations of both GO and UMLS knowledge bases.

Exact String Matching. The first and simplest method was exact string matching, which finds the lexical matches between UMLS terms and GO terms. This was implemented through a join of MySQL tables containing the set of terms from both the GO and the UMLS.

Norm. Norm represents one of two methods utilized from the lexical tools available from the UMLS.[13] As its name implies, *norm* converts text strings into a normalized form, removing punctuation, capitalization, stop words, and genitive markers. Following the normalization process the remaining words are sorted in alphabetical order. After processing each GO term with *norm,* we matched each distinct GO term against the normalized form of UMLS terms, as represented by a normalized English table that was obtained from the NLM (MRXNS_ENG).

MMTx. MMTtx is an implementation of MetaMap, a highly configurable program used to map concepts found in biomedical texts to the UMLS Metathesaurus.[22] After normalizing and parsing text into noun phrases, MetaMap uses UMLS knowledge sources to generate both synonyms and derivational variants.

The program then retrieves the set of all candidate strings contained in the Metathesaurus, evaluates each one against the input text, and produces a result list of candidates and their UMLS CUIs, ordered by mapping strength (scored on a scale from 0-1000). The highest-scoring combinations of the candidates are presented as an additional list called 'Meta Mappings' that also scores the results.

We used MMTx to perform two types of analyses, which we term 'Loose' and 'Strict'. Loose MMTx analysis consists of all distinct GOID-CUI pairs provided by the 'Meta-Mappings' regardless of their scores. For 'Strict' analysis, we determined a threshold score (T) and neglected all mappings below the selected threshold. We varied the thresholds from T=600 to T=1000, in increments of 100, to provide five different sets GOUI-CUI pairs.

BLAST-based Matching. Based on the work of Krauthammer, et al., the Basic Local Alignment Search Tool (BLAST)[24] can be used to compare text characters to each other that have been converted into nucleotide sequences.[25].BLAST allows for approximate string matching, with the respective flat files serving as reference databases. BLAST provides efficient identification of sequences that have a high probability of correspondence with the query sequence. We transcribed the text strings of both the GO and the UMLS terms into nucleotide sequences by substituting each character with a specified nucleotide combination, as outlined by Krauthammer, et al[25]. After converting the transcribed nucleotide sequences into FASTA format, we created a BLAST database using the *formatdb* program of the transcribed UMLS terms. Using the *blastn* application, each transcribed GO term was used as a query sequence to search the entire UMLS database for the most similar match.

3.2 Implementation

Every mapping methodology was used to map a subset of GO terms provided by the NCI to the entire UMLS. While GO has only one term per concept, UMLS supports multiple variant terms for every concept. This results in a potential total of 574 million pairs (332 GO terms * 1,728,075 SUIs) that need to be searched. Since we are only concerned with mapping the *concepts* (as opposed to *terms*) of the respective terminologies, this number is further reduced since we consider only the possible *distinct* GOID-CUI pairs stemming out of the mapping methods (this is a smaller subset since many UMLS terms can be represented by a single CUI). The resulting concept combinational space consists of 265 million individual GOID-CUI pairs (332*797,359).

3.3 Evaluation

The GS, consisting of 332 GO unique identifier/UMLS Concept Unique Identifier pairs, was examined for each method and analyzed in the following manner: relevant

pairs ('True Positive'; TP) were pairs found by the coupling method that were also in the GS; non-relevant ('False Positive'; FP) matches were those that were not found in the GS; relevant, but *not* retrieved ('False Negative'; FN) were in the original GS but not matched by the coupling method. Using the GS GOID-CUI pairs as our gold standard, we performed an evaluation of each of the methodologies implemented.

We measured the efficacy of each of the methods using precision and recall. Recall was calculated as the ratio of the number of *distinct* GOID-CUI pairs that were identified by the mapping method that *match* GOID-CUI pairs in the GS, divided by the total number of pairs in the GS, TP/(TP+FN). Precision was measured as the ratio of the number of *distinct* GOID-CUI pairs returned by the mapping method that *match* GOID-CUI pairs in the Gold Standard, divided by the total number of putative GOID-CUI pairs found by the mapping method, TP/(TP+FP).

4 Results and Discussion

Table 1 and figure 1 summarize the analysis of our results using each of the four described coupling methodologies.

Within the context of the gold standard, our evaluation shows that each of the four coupling methods provides mapping with varying degrees of success. The simpler coupling approaches, exact string match and *norm*, fared well in terms of precision, but suffered in recall. Exact match, the simplest of the methods tested, had a recall of only 65% but had a high level of precision (94%). Because of what constitutes an exact match, this high level of precision is to be expected. On the other hand, *Norm* yielded both a high recall (90%) and precision (89%). The 4-5% rate of non-relevant matches (FP) for these relatively simple matching methods is reflective of the variance, and hence the difficulty in finding all the match pairs, that exists in the UMLS of a single concept's terms. In an ideal world, if everyone used the same terms for the same concepts, then exact string algorithms would yield 100% precision.

The more complex methods, BLAST and Loose-MMTx, both suffered in precision (36% and 62% respectively) as a result of how their general algorithms work and how the results were tabulated. That is, both of these methods attempt to make a classification for almost every term that it is given resulting in as many FP as FN counts (Loose-MMTx recall = 82%; BLAST recall = 36%). Imposing the strictest threshold (T=1000), Strict-MMTx achieved a precision of 95%. However, recall dropped to 74%. We explored changing the threshold to various values, and report the values at threshold equal to 900 and 800 graphically in figure 1. There was a consistent trade-off between precision and recall as we changed the threshold

value. As there was no significant effect for thresholds tested below 800, these values are not shown in the graph.

Table 1. Analysis of GO and UMLS Mappings according to the methodologies (Count of Distinct GOI-CUI Pairs)

	Exact String	Norm	MMTx Loose	MMTx T=1000	BLAST
Relevant GS Matches (TP)	216	299	272	247	121
Non-Relevant GS Matches (FP)	13	38	170	12	211
Not Retrieved GS Matches (FN)	116	33	60	85	211

BLAST's low precision may be due in part to how the BLAST algorithm functions and highlights some of the issues of the heuristic method that BLAST implements. BLAST is designed to try to determine the best alignment between two nucleotide sequences, a 'query' and 'subject' sequence, based on contiguous word lengths as defined by the algorithm. Certain penalties are incurred for the incorporation of a gap (or space) that may be inserted into a sequence to help optimize and alignment. In the present study, no gaps were inserted in any sequences. Possible lowering of the incurred penalty for inserting a gap may have yielded a higher recall value.

It is important to note that the alignment that is sought by BLAST is for the longest sequence with the least amount of dissimilarity. This can be problematic in implementations where slight variants of words result in the transcribed sequences being different enough to not make them look similar. Also it is important to point out that BLAST is an algorithm of *similarity*, not *homology*, which can make the heuristic prone to misclassifications. For example, BLAST addresses the question: "Which item(s) in the database are *most similar* to my query sequence?" This is opposed to addressing the question: "Which item(s) in the database *is the same as* my query sequence?"

As a result, for every sequence an attempt is made by the algorithm to find a match. This produces not only a low recall value, but (by definition) results in a high number of irrelevant matches (64%). Similar issues have been addressed in relation to attempting to use BLAST as a primary classification tool for nomenclature of traditional biological sequences.[26]

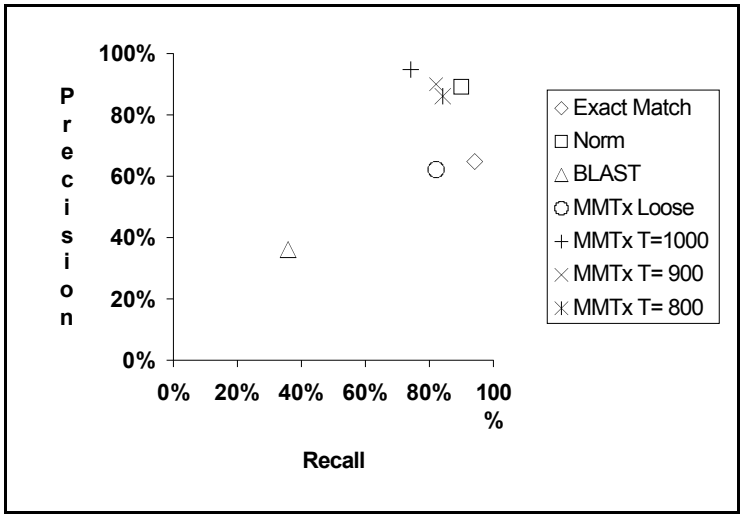

Figure 1: Precision versus Recall according to the mapping method

As the gold standard file used for the present study consisted of a heavily curated data set that is maintained by the National Cancer Institute, our results may have been biased, since it represented a subset of the UMLS, particular to only one domain of medicine (oncology). As a result, specific UMLS concepts that may not be contained in the NCI metathesaurus may not have been mapped to a GOID. Additionally, because of the millions of SUIs that a single GO term could map to, it is unreasonable to expect manual curation of all combinations. This may contribute to an inflated number of calculated False Positive values. To this end, we qualitatively explored some of the False Positives reported by the mapping methods.

Two experts reviewed some of the distinct GOID-CUI pairs that were classified as False Positive. They unveiled that some GOID-CUI pairings labeled as FP, because they did not exist in our GS, appeared to be relevant. For *exact string match* and *norm*, 13 and 8 of their respective GOID-CUI FP matches appeared correctly mapped and thus misclassified as FPs. The recalculated precision of the two methods would reach 100% and 91%, respectively. Similarly, some of the irrelevant matches returned by both of MMTx and BLAST also appeared misclassified, but their large sets of FP counts precluded quantitative analyses. This issue is reflective of the complex task of mapping terms to a composite terminology like the UMLS, which may have multiple synonymous terms and contains millions of distinct terms.

This last point emphasizes the complexity of the issue that is at the core of ontologies for use in a practical setting. In order to address the non-homogeneous manner in which professionals choose to represent the same concept a number of

approaches have been implemented. The Systematized Nomenclature of Medicine (SNOMED) is an example of a terminology that has taken this factor into account and allows concepts to be defined in terms of other concepts using formal predicate logic[27]. This formalism allows for computable evaluations of incomplete 'identity' matches in which one of the following relationships could be assigned: 'putative ancestor', 'putative descendant' or 'putative sibling'.

These results demonstrate the difficulty and applicability of existing coupling methods for mapping even a very select set of terms between terminologies. The difficulty of mapping arises, in part, because the UMLS is a composite terminology. That is, it is made of not just a single resource, but multiple resources. To this end, UMLS inherits all the terms from all the various terminologies that it includes. For this reason, NLM has developed a number of tools, such as *norm* and MMTx to map novel terms to the existing unique concept identifiers. To date, the tools from NLM have been solely used to append additional medical terminologies. While some gene and gene product terms may have been incorporated from existing medical terminologies there has been no attempt, to our knowledge, for adding strictly basic biological terms.

Appending basic biological terminologies to the UMLS is a different art than adding terms from medical terminologies. Processing text strings with tools such as MMTx or *norm,* for example, de-emphasizes numeric values, which are less important in clinical medicine (though there are exceptions, such as 'Diabetes Mellitus Type 2') than in biology, where such differences are extremely important. Basic biological terms often consist of subtle textual, yet significant definitional differences, such as the difference between 'Transcription Factor I', 'Transcription Factor II', or 'Transcription Factor III'. Using CUIs or GOIDs provides a method for overcoming these types of ambiguities, since appropriate concept mapping is essential for proper integration of the respective terminologies.

5 Caveats and Implications for Future Work

It is important to emphasize that the gold standard used in the present study was only a relatively small, domain-specific subset based on work that was not specifically aimed at mapping GO terms to the UMLS. Since creating a gold standard is a resource intensive endeavor, we conveniently used relevant work from the NCI. However, as future work will inevitably include GO in the UMLS, it will be imperative to be aware of such mappings, especially in the case of certain concept terms that may be described very differently by either clinicians or biologists. These will likely involve either manual curation or sophisticated Natural Language Processing (NLP) tools to create mappings.

Nonetheless, this work demonstrates the feasibility of mapping a significant portion of GO to existing UMLS concepts. It is conceivable that using a combination of the methods described in the present study, as well as others that may exist, will enable a complete mapping of these terminologies as well as any future terminology from other related domains. We are conducting additional work that will address the insertion of concepts in the UMLS schema with algorithms allowing the mapping of terminologies with additional relationships beyond the "identity" relationship, such as the "ancestor", "descendant" and "sibling" relationships.

Beyond examining additional methods for simple mapping terms, integration of these techniques into a voting scheme may also yield more favorable results by incorporating the strengths of each of the methods used. While the present study primarily focused on the ability of each of the methods to function independently, we believe that by linking the methods together in a simple voting scheme may result in higher levels of precision and recall for the overall mapping process.

6 Conclusions

This work lays the foundation for mapping biological and medical terminologies, which are essential for information systems. The task of knowledge mapping will be essential for the future integration of the various biomedical domains, such as nursing, public health, etc. The primary challenge that these future mapping face, as is faced with GO, is that clinical informaticians may not necessarily design a significant portion of these non-medical terminologies, leading to potential ambiguities, redundancies, or incompatibilities Using proven text mining and information extraction techniques it will be possible to create a mapping of large sections of existing knowledge resources to each other.

Acknowledgments

INS and MNC are funded by National Library of Medicine Medical Informatics Training Grant LM07079-09. Additional financial support for this work was from LM06274 of the National Library of Medicine. Special thanks goes to Sherri De Coronado of the National Cancer Institute for her assistance and providing us with their UMLS and GO mappings. Thanks also goes to David Figurski and his laboratory for their encouragement and insightful discussions, without which this paper would not be possible.

References

1. Altman RB & Klein TE. Challenges for Biomedical Informatics and Pharmacogenomics. Annual Review of Pharmacology & Toxicology. 42:113-133. (2002)
2. Nakamura RM. Technology That Will Initiate Future Revolutionary Changes in Healthcare and the Clinical Laboratory. J Clin Lab Anal. 13:49-52. (1999)
3. Shortliffe EH & Perrault LE (eds). Medical Informatics: Computer Applications in Health Care and Biomedicine. (Springer, New York, 2001)
4. Lindberg DA, Humphries BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 32(4):281-291. (1993)
5. Yu H, Friedman C, Rhzetsky A, Kra P. Representing Genomic Knowledge in the UMLS Semantic Network. 1999 Proc Annu Symp Am Med Inf Assoc:181-186. (1999)
6. Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. Gen Res. 11(8):1425-1433. (2001)
7. Bodenreider O & McCray AT. The Lexical Properties of the Gene Ontology. Proceedings of the American Medical Informatics Association 2002 Annual Symposium.
8. Cimino JJ & Johnson SB. From ICD-9 to MeSH Using the UMLS: A How-to Guide. 1994 Proc Annu Sump Am Med Inf Assoc: 730-734.
9. Zeng, Q & Cimino JJ. Mapping Medical Vocabularies to the Unified Medical Language System. 1996 Proc Annu Symp Am Med Inf Assoc: 105-109. (1996)
10. Tuttle MS, Suarez-Munist ON, Olsen NE, et al. Merging Terminologies. 1995 MEDINFO. 8(Pt 1):162-166. (1995)
11. Tuttle MS, Cole WG, Sheretz, DD, Nelson SJ. Navigating to Knowledge. Methods Inf Med. 34(1-2):214-231. (1995)
12. Tuttle MS, Sherertz DD, Erlbaum MS, et al. Adding Your Terms and Relationships to the UMLS Metathesaurus. 1991 Proc Annu Symp Comput Appl Med Care:219-223. (1991)
13. National Library of Medicine. UMLS Lexical Tools. Application and Documentation available at http://umlsks.nlm.nih.gov.
14. Lussier YA, Shagina L, Friedman C. Automating SNOMED Coding Using Medical Language Understanding: A Feasibility Study. 2001 Proc Annu Symp Am Med Inf Assoc: 418-422. (2001)
15. McCray AT, Bodenreider O, Malley JD, Browne AC. Evaluating UMLS Strings for Natural Language Processing. 2001 Proc Annu Symp Am Med Inf Assoc: 448-452. (2001)
16. Friedman C, Liu H, Shagina L, et al. Evaluating UMLS as a Source of Lexical Knowledge for Medical Language Processing. 2001 Proc Annu Symp Am Med Inf Assoc: 189-193. (2001)

17. Raychaudhuri S, Chang JT, Sutphin PD, Altman RB. Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature. Gen Res. 12:203-214. (2002)
18. Sperzel WD, Abarbanel RM, Nelson SJ, et al. Biomedical Database Inter-connectivity: An Experiment Linking MIM, GENBANK, and META-1 via MEDLINE. 1991 Proc Annu Symp Comput Appl Med Care:190-193. (1991)
19. National Library of Medicine. Unified Medical Language System. 12th Ed. January 2001.
20. Tuttle MS, Olsen NE, Campbell KE, et al. Formal Properties of the Metathesaurus. 1994 Proc Annu Symp Comput Appl Med Care:500-504. (1994)
21. Coletti MH & Bleich HL. Medical Subject Headings Used to Search the Biomedical Literature. J Am Med Inf Assoc. 8(4):317-323. (2001)
22. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc. 88(3):265-266. (2000)
23. Aronson AR. Effective Mapping of Biomedical Text to the UMLS. 2001 Proc Annu Symp Am Med Inf Assoc: 17-21. (2001)
24. Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. J Mol Biol. 215(3):403-410. (1990)
25. Krauthammer M, Rzhetsky A, Morosov P, Friedman C. Using BLAST for Identifying Gene and Protein Names in Journal Articles. Gene. 259(1-2):245-252. (2000)
26. Sarkar IN, Thornton J, Planet PJ, et al. An Automated Phylogenetic Key for Classifying Homeoboxes. Mol Phylogenet Evol: 24:388-399. (2002)
27. Spackman KA & Campbell KE. Compositional Concept Representation using SNOMED: Towards Further Convergence of Clinical Terminologies. 1998 Proc Annu Symp Am Med Inf Assoc: 875-879. (1998)