Exploring the Role of Knowledge Representation and Reasoning in Biomedical Text Understanding

Debra T. Burhans
Canisius College
2001 Main Street
Buffalo, NY 14208
1.716.888.2433
burhansd@canisius.edu

Alistair E. R. Campbell
Hamilton College
198 College Hill Road
Clinton, NY 13323 |
1.315.859.4377
acampbel@hamilton.edu

Gary R. Skuse
Rochester Institute of Technology
85 Lomb Memorial Drive
Rochester, New York 14623
1.585.475.2532
grssbi@rit.edu

ABSTRACT

There is considerable effort being devoted to mining information from medical and scientific literature, in particular, from Medline abstracts and from full-text articles. Such information is being used, for example, to reconstruct biological pathways, identify pathogenic mechanisms and, importantly, to identify functional relationships that can be used to predict disease onset and its course thereafter. Our interest is in exploring the role of knowledge representation and reasoning (KR&R) as it relates to the problem of understanding biomedical text. The role we envision for a KR&R system in this context is as a knowledge store for a small, focused subset of abstracts gleaned from the Medline corpus that is relevant to a problem of interest. The system will infer new knowledge from the represented abstracts which can then be stored in a larger data repository and reported to a biologist. We are specifically interested in designing a system that, given a set of abstracts, can perform many of the same inferences that a biology expert would make if given the same set of abstracts. Inferences that go beyond the predictions of the biologist are particularly interesting, but our initial goal is to emulate the biologist. We have selected the disease neurofibromatosis type 1 (NF1) for our study with the goal of developing a model that can be applied to reasoning about other diseases and problems of interest. This approach is focused narrowly on a particular problem and as such may not lead to solutions relevant for general problem solving. However, we believe there is an important role for specialized problem solvers in the larger context of biomedical text understanding. Working closely with a domain expert in biology is providing valuable insights into how the computational synthesis of information might best serve the needs of a biologist. This preliminary report describes our current work on hand analysis and translation of abstracts and our proposed overall approach to the problem.

Authors' addresses: Debra T. Burhans, Computer Science Department, Canisius College, NY 14208, Alistair E.R. Campbell, Department of Computer Science, Hamilton College, 13323 , Gary R. Skuse, Department of Biological Sciences, Rochester Institute of Technology, 14623.

1. INTRODUCTION

Considerable effort is being devoted to mining information from medical and scientific literature, in particular, from Medline abstracts and from full-text articles. Such information is being used, for example, to reconstruct biological pathways, identify pathogenic mechanisms and, importantly, to identify functional relationships that can be used to predict disease onset and its course thereafter.

Our interest is in exploring the role of knowledge representation and reasoning (KR&R) as it relates to the problem of understanding biomedical text. Rather than work on the development of information extraction tools that can be applied to large corpora, such as the Medline abstracts, we are interested in understanding how one biologist reasons about information contained in biomedical literature related to his area of expertise, in this case, neurofibromatosis type 1 (NF1). This small, exploratory project is focused on building a KR&R system that is capable of inferring new information that corresponds to that inferred by the biologist. An underlying assumption of this approach is that there is significant overlap among biologists who work in similar areas in terms of what they will infer given a set of abstracts. We also assume that there are generalities that can be captured using this approach that will allow us to use this work as a model for other biomedical problems. Working closely with a domain expert in biology is providing valuable insights into how the computational synthesis of information might best serve the needs of a biologist.

The role we envision for a KR&R system in this context is as a knowledge store for a small, focused subset of abstracts gleaned from the Medline corpus that is relevant to a problem of interest. This project is complementary to large-scale information extraction efforts, and depends upon those efforts to extract the concepts and relations to be represented in our KR&R system.

Herein we report preliminary progress. Section 2 describes the background against which the project is set and gives details about the proposed KR&R framework to be used. Section 3 presents our overall approach to the problem of constructing a biology reasoning agent. Section 4 describes our current work on abstract analysis. Section 5 presents some of our knowledge structures and a discussion of issues encountered while developing knowledge representation that will support the inferences necessary for the system we envision. Section 6

concludes with challenges and future directions for this project and section 7 lists the references cited.

2. BACKGROUND

2.1 Biomedical Text Mining

There is considerable effort being devoted to mining information from medical and scientific literature, in particular, from Medline abstracts and full-text articles [4,5,8,9,10,15,26,29]. Such information is being used to reconstruct biological pathways, help link related research articles with empirical work and to make discoveries *in silico*. We will make use of the tools developed by others in order to allow us to extract the relations and concepts necessary to construct the knowledge representation.

2.2 SNePS

The KR&R system we are using is SNePS [27], a powerful and extensible knowledge representation and reasoning system. A SNePS network can be viewed as comprising the mind of a cognitive agent. SNePS has been used to build cognitive agents in a variety of areas, including cognitive robotics and natural language understanding. SNePS is a general-purpose knowledge representation and reasoning system that pays particular attention to the representation of propositions as nodes in a semantic network. It is particularly well-suited to the task of reasoning about biological information due to the use of contexts.

Conceptually, a context is a virtual belief space. SNePS contexts are a feature that supports reasoning over multiple sources of knowledge, which is clearly important in this domain. In particular, the representation of information from different sources in different contexts allows a proposition and its negation to coexist across contexts without causing a contradiction. Indeed, entirely separate and contradictory theories may be maintained in the knowledge representation through the use of contexts. Furthermore, if a single context contains contradictory asserted propositions, SNePS gracefully handles these without deriving spurious unsupported propositions and without inhibiting continued inference with the non-contradictory information in the context. This will be invaluable due to the plethora of information sources that must be integrated.

Reasoning may proceed with any subset of the contexts available to the reasoning agent. A further strength of contexts involves the usefulness of the proposed system to individual scientists. A context, or set of contexts, can be used to represent any information, including that which is unique to a particular scientist or laboratory. Thus, an individual's beliefs, hypotheses, and data can be represented in an appropriately designed set of contexts and included (or excluded, as desired) from reasoning. This feature enables individual scientists to test and verify hypotheses in addition to demonstrating the ramifications of their experimental data with respect to what has already been published.

2.3 Medline Abstracts

Abstracts have been chosen as the target of this work for several reasons. First, the Medline database is available, curated and updated regularly. Second, full text articles are not uniformly available. Some are available without a subscription to the publishing journal, some require subscriptions, and others are simply not available under any circumstances. When available, full-text articles are published in a variety of formats including ASCII text and Adobe portable document format (i.e. pdf). This variability complicates processing of those articles. Third, our hypothesis is that most of the information needed for reasoning is contained in the abstract. Finally, there are freely available tools being constructed to parse Medline abstracts [4,5,8,9,10,15,26,29].

2.4 Neurofibromatosis Type 1 (NF1)

The choice of NF1 was based on the expertise of the biologist in our research group as well as a number of features of NF1 related to biological and medical considerations. NF1, or von Recklinghausen disease, is the most common hereditary disease predisposing to cancer [17]. In addition to malignant tumors, affected individuals develop benign tumors (i.e. neurofibromas, the hallmark of this disease), skeletal abnormalities, café au lait spots, Lisch nodules of the iris and learning disabilities [13,20]. Despite the fact that the NF1 gene was isolated more than one decade ago, there has been little progress toward associating specific genetic abnormalities with specific clinical manifestations. Indeed only one region, comprising approximately 1500 of 11,000 nucleotides in the most common of several alternative transcripts, has been associated with any function [1]. NF1 has complex clinical manifestations and is incompletely understood. Despite the countless clinical and laboratory studies published over the past several decades there remains no consensus regarding genetic mutations and clinical features [30]. The work described herein may serve to uncover relationships heretofore unknown.

3. APPROACHES TO THE PROBLEM

Our initial approach is to select and represent a limited set of Medline abstracts related to NF1. This involves building up a representation that enables the types of inference that a biologist with expertise in NF1 makes when reading and synthesizing abstract texts on NF1.

3.1 Selection and Representation of Abstracts

Specifically, our goals are to:

1. Select an initial set of 100 abstracts on NF1 for system development and training. Those abstracts will be selected from literature published within the past five years and summarized by hand.
2. Biologically significant inferences that can be drawn from these subsets will be noted. Those inferences correspond to information not explicitly represented in any of the abstracts, but inferable by integrating information across abstracts or within a particular abstract. This interpretation will guide the development and modification of the representation and reasoning rules in order to enable the system to make at least half as many inferences as can be drawn by hand.
3. Develop SNePS representations for these abstracts, creating new case frames and reasoning rules as the need arises. Case frames are relations between concepts. These relations will be based on current work in ontologies and knowledge representation of biology data [2,6,19,21,22,28].

At this point we have selected and partially analyzed a set of 20 abstracts described in the next section. We have also started work on a case frame dictionary for the project, creating new case frames in order to support the types of reasoning required to infer biologically significant information.

3.2 Ontology Resources

The development of new case frames, in addition to the language understanding issues, has led us to employ several ontology resources in this project. The suggested upper model ontology (SUMO) is a freely-available, provably consistent set of terms and definitional axioms that describe the most basic concepts of commonsense knowledge [19]. The Unified Medical Language System (UMLS) is a unified collection of source vocabularies concerning the domain of medicine (http://nlm.nih.gov/research/umls). A particularly useful piece of the UMLS is the Semantic Network, a set of semantic concepts and relations that form a directed graph in which the semantic concepts are the nodes and the relations are the arcs. We have developed a translation tool that encodes the UMLS Semantic Network into SNePS. The Semantic Network can be considered a kind of ontology; however the concepts at the upper level are not as well-considered as in SUMO, and the relations are not as precisely defined using formal logic as they are in SUMO. We propose to merge the UMLS Semantic Network with the SUMO upper ontology into a single general-purpose biomedical ontology.

The WordNet on-line lexical database organizes concepts into several hierarchies, where each concept is represented by a synonym set (synset) of terms that synonymously denote that concept [3]. Comprising over 100,000 different English words, WordNet is the largest and most widely used lexical database. Recent work in ontology merging has produced a mapping of every word sense in WordNet onto SUMO concepts [18].

3.3 Context

While we have chosen to focus on a single disease, there are still problems distinguishing relevant from non-relevant abstracts. Neurofibromatosis type II is a related but distinct disease, and abstracts about NFII may be difficult to filter out in our literature searches. In addition, nuclear factor 1 is also referred to as NF1 in the literature, meaning that we have to disambiguate between the two abbreviations by using contextual information. Finally, there are papers that discuss both neurofibromatosis 1 and nuclear factor 1 together that are of interest, so simply filtering out papers on nuclear factor 1 is not a solution to this problem.

3.4 System Evaluation

The hand selection, representation, and evaluation of abstracts and inference performed using the information therein comprises a training phase that will help to elucidate which information from the abstracts is required in order to infer new, interesting information as well as to demonstrate the sort of background information that is necessary in order to perform reasoning. This process will be reiterated until the system is performing reasonably well, where our initial criterion for this is that the system is capable of inferring at least 50% of what is inferred by the biology expert. Once the system is performing automated reasoning successfully as defined above, a second test set of 100 abstracts will be selected and represented. The inferences derived from these abstracts by the system will be compared to those derived by the biology expert. The overlap of these two sets must include at least 50% of the biologist's inferences in order for the experiment to be considered successful.

4. ABSTRACT ANALYSIS

A set of 20 abstracts has been selected for initial analysis. 10 non-overlapping pairs of abstracts were created at random and information not explicitly represented in either abstract in each pair but inferable from the pair was noted by the biologist. In addition to these conclusions, the biologist was asked to summarize each abstract and to highlight those portions of the abstract text he used in drawing his conclusions.

4.1 An Abstract Pair

The first abstract selected (A1) is from a paper by Reish et al. entitled "Modified allelic replication in lymphocytes of patients with neurofibromatosis" [24]. The second abstract in this pair (A2) is from a paper by Okazaki et al. entitled "The mechanism of epidermal hyperpigmentation in café-au-lait macules of neurofibromatosis type 1 (von Recklinghausen's disease) may be associated with dermal fibroblast-derived stem cell factor and hepatocyte growth factor" [23].

A1 starts with a general statement, "Transcription activity of genes is related to their replication timing." It goes on to discuss a perturbation in replication timing that is associated with NF1 patients but not necessarily with malignancies. There is nothing in A1 to indicate whether the perturbation affects the expression levels of the implicated genes (they could be expressed at the same level or over- or under-expressed). The phrase "related to" implies a bidirectional relation, in this case between transcription activity and replication timing. This relation is underspecified, which is not unusual in biomedical text. What we may conclude, regardless of the way in which two processes are related, is that when one is affected the other one ia also affected: if a proposition is true of one then there is a proposition (not necessarily the same proposition) that is true of the other. The conclusion in this abstract used by the biologist to help infer new information, along with A2, is the fact that NF1 is associated with the activation of cancer-implicated genes.

A2 indicates that there is an excess of two growth factors in NF1 patients. From A1 and A2 together, the biologist inferred that, since perturbations in cancer-implicated genes are associated with NF1, and HGF and SCF are abnormally expressed in NF1 patients, HGF and SCF may be associated with cancers.

4.2 Observations

We have noted the following in analyzing the notes generated by the biologist:

1. The conclusions drawn by the biologist thus far have involved approximately one-tenth of the text in the abstracts. In some cases the abstract title alone was sufficient to draw interesting conclusions. This seems to indicate the possibility in the future of automatically ignoring redundant or background information in abstracts thereby reducing the complexity of the representation task.

2. A number of important pieces of background knowledge employed thus far by the biologist are captured in the ontology resources discussed in the previous section.

3. Abductive reasoning is pervasive and important. This has important implications for the system selected to reason about the abstracts. In particular, it suggests that database models, and specifically SQL queries, will not be sufficient for performing inference.

5. KNOWLEDGE REPRESENTATION

The initial knowledge representation task for this project is to develop case frames for representing biomedical information. A case frame in SNePS is a set of relations (arcs) among nodes. For example, the SUBCLASS-SUPERCLASS case frame comprises a node that represents the proposition that A is a SUBCLASS of B with a SUBCLASS arc from this node to node A and a SUPERCLASS arc from this node to node B. Developing knowledge representation for biological information is a true interdisciplinary effort and challenge. Abstracts contain information at a variety of levels of detail and abstraction. In order to facilitate future reasoning and translation, in particular, that which is biologically meaningful, case frames must reflect the underlying science. For example, the *ras* protein is important in both enhancement and suppression of carcinogenesis. These are "contradictory" processes. In addition *ras* is not a single molecule but a collection of many molecules. In the future it may be possible to describe experimental data at the level of individual molecules. In developing a representation for proteins such as *ras* we need to allow for the possibility of changing the granularity of the representation without disrupting the already-existing knowledge base. We also need to account for the fact that, under different circumstances, proteins behave differently and may have contradictory functions.

The notion of a "biological context", defined here as a collection of physical entities, is critical to understanding and representing information about the state of a particular set of cellular processes of interest. The literature in the biomedical field often refers to changes of state, for example, changes in what is bound by a protein. When *ras* binds GDP (guanine diphosphate) it is inactive, but when it binds GTP (guanine triphosphate) it is activated. This change is important when production of the NF1 protein is suppressed. Each of these states, in conjunction with other relevant information, would be represented in a different biological context. Biological contexts in turn exhibit a partial temporal ordering.

The notion of a biological context differs from what is termed a "propositional context" in SNePS. A propositional context is generally a set of consistent propositions, though SNePS permits the representation of contradictory information without compromising logical inference by allowing a user to indicate which of the contradictory propositions (if any) should be used for reasoning. We expect to associate a single abstract with its own propositional context. Within such a context we need to be able to represent and reason about more than one *biological* context for a particular biological entity, for example, a protein. This means that at least some of the information for a biological entity must be identified with a particular biological context since the properties of the entity in different contexts may be contradictory. The need to tag propositions with this type of information leads to the classic three-dimensional/four-dimensional choice for modeling the universe, and it is generally unresolved. The three-dimensional approach tags each proposition with a context arc, resulting in a plethora of arcs. The four-dimensional approach uses a different intensional representation for each temporal abstraction, in this case, for each biological context, resulting in a plethora of nodes. Given that the primary inference in SNePS is node-based, rather than path-based (arcs), it will be more efficient to minimize the number of nodes and use a three-dimensional model.

5.1 Example of Representation

The following example demonstrates some of our representational structures. The sentence to be represented is one of a number that comprise background knowledge needed to successfully understand and reason about a particular abstract. The phrases were generated by hand and involved careful examination of the abstract and consultation between a computer scientist and biologist.

The sentence is, "In every context *ras* binds GTP or GDP but not both." Given the foregoing discussion of contexts, this will be represented as a universally quantified rule that matches every occurrence of *ras* in every biological context. Every ras in every context is the rule antecedent (`&ant` indicates conjunction in the antecedent), and is represented as follows:

```
(assert forall ($con $r)
  &ant (build member *con class
         context)
  &ant (build member *r class ras)
  &ant (build member *r collection
         *con)
```

When a variable is first introduced in a rule it has a dollar sign as a prefix, after which it is prefixed with an asterisk. `con` is the context and `r` is the *ras*. The third conjunct represents that *ras* is a member of a *collection*, which in this case is the context.

The consequent of the rule (`cq`) begins:

```
cq (build min 1 max 1
```

This indicates that exactly one of the arguments in the consequent is true (`min 1` means a minimum of 1, `max 1` means a maximum of 1). The two arguments are "bound to GDP" and "bound to GTP". The structure for the GTP argument is shown (GDP is analogous.)

```
arg ((build bound *r
 bound (= (build skf gtp_for *con) m1))
         (build member *m1 class "GTP")
         (build member *m1 collection
             *con)))
```

`bound` is defined as a Skolem function (`skf`) that indicates that some GTP that depends on the context under consideration is bound to the *ras* in that context. `m1` is a node that represents the proposition that this is the case.

5.2 Discussion

The abstract that led to the creation of the initial set of background information phrases is entitled, "The prognostic significance of bone marrow levels of neurofibromatosis-1 protein and ras oncogene mutations in patients with acute myeloid leukemia and myelodysplastic syndrome" [16]. From this abstract four phrases were identified as containing critical information for reasoning. Thirty-seven phrases containing background information that would enable the understanding of these four phrases were written. The background information includes synonymy, ontological relations and general facts about biology. All of this information may be available from existing knowledge resources. The identification of the types of background information needed for abstract understanding is an important step that will inform the selection of these resources as well as the creation of new resources as necessary.

Translating abstracts into knowledge representation by hand is clearly a labor-intensive task. However, the background knowledge identified thus far relates to all of the abstracts under consideration. We expect that after our initial hand-translation phase there will be little additional background information required to understand new abstracts.

6. CHALLENGES AND FUTURE WORK

Hand-annotating and representing abstracts are extremely time-consuming tasks and subject to error. This initial phase will help us determined the efficacy of the overall approach to reasoning with information derived from biomedical texts. Future work clearly will require automatic acquisition of information from abstracts. Specifically, we plan to carry out the following tasks:

1. Scaling up the system (automating the translation from text into XML and XML into SNePS).

2. Building an interface between SNePS and the DB2 (http://www-3.ibm.com/software/data/db2/) database management system (where we currently store Medline data). Note that this involves developing a protocol for determining when storage of newly inferred information should be undertaken (from SNePS to DB2), and when a piece of information should be removed from the "current memory" of the reasoning agent (SNePS concepts).

In summary, we are interested in exploring the utility of a employing a powerful KR&R system to reason about biomedical texts. Our initial approach to the problem involves hand-translation and annotation of abstracts, which we recognize is untenable in the long run and must be replaced with automatic translation. We have formed a true interdisciplinary team in order to build a system that embodies the underlying biological concepts. The inclusion of a biologist is helping to elucidate the types of inference that may be important to biologists and the background information used in performing those inferences.

Our hypothesis is that there is something to be gained by applying a powerful reasoning system to information gleaned from Medline abstracts, and that doing so will allow us to partially emulate the reasoning processes of a biologist. This should advance the goals of computational approaches to information extraction in this field, in particular, to aid biologists in designing experiments and to help them verify or refute potentially important hypotheses.

7. REFERENCES

[1] Cichowski, K. and Jacks, T. (2001) NF1 tumor suppressor gene function: narrowing the GAP. Cell 104, 593-604.

[2] Colomb, R.M. and Weber, R. (1998) Completeness and quality of an ontology for an information system. Proc. Formal Ontology in Info. Sys. 207-217.

[3] Fellbaum, C. ed. *WordNet: An Electronic Lexical Database* The MIT Press, 1998.

[4] Friedman, C., Kra, P., Yu, H., Krauthammer, M. and Rzhetsky, A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17, 574-582.

[5] Grover, C. Klein, E. Lapata, M. and Lascarides, A. (2002) XML-based NLP tools for analyzing and annotating medical language. Proc. Second Int. Workshop on NLP and XML.

[6] Guarino, N. and Welty, C. (2000) A formal ontology of properties. Knowledge Eng. And Knowledge Management Meth., Mod. And Tools 12th Int. Conf. EKAW2000, 97-112.

[7] Hahn, U. and Schulz, S (2003) Towards a Broad-Coverage Biomedical Ontology Based on Description Logics, Pacific Symposium on Biocomputing 8, 577-588.

[8] Hanisch, D., Fluck, J. and Mevissen, H.-T. (2003) Playing biology's name game: identifying protein names in scientific text. Pacific Symposium on Biocomputing 8, 403-414.

[9] Hatzivassiloglou, V., Duboue, P.A. and Rzhetsky, A. (2001) Disambiguating proteins, genes, and RNA in text: A machine learning approach. Bioinformatics, (ISMB Supplement) 1-10.

[10] Hirschman, L., Park, J.C., Tsujii, J., Wong, L. and Wu, C.H. (2002) Accomplishments and challenges in literature data mining for biology. Bioinformatics 18, 1553-1561.

[11] Karp, P.D., Ouzounis, C. and Paley, S. (1996) HinCyc: A knowledge base of the complete genome and metabolic pathways of *H. influenzae*. Proc. of Intel. Sys. Mol. Biol. 1-15.

[12] Kohler, J. and Schultze-Kremer S. (2002) The semantic metadatabase SEMEDA): ontology based integration of federated molecular biological data sources. In Silico Biol. 2, 19-31.

[13] Korf, B.R. (2000) Malignancy in neurofibromatosis type 1. Oncologist 5, 477-485.

[14] Lambrix, P. and Edberg, A. (2003) Evaluation of Ontology Merging Tools in Bioinformatics, Pacific Symposium on Biocomputing 8:589-600.

[15] Leroy, G. and Chen, H. (2002) Filling preposition-based templates to capture information from medical abstracts. Pacific Symposium on Biocomputing 7, 350-361.

[16] Lu, D., Nounou, R., Beran, M., Estey, E., Manshouri, T., Kantarjian, H., Keating, M.J. and Albitar, M. (2003) The prognostic significance of bone marrow levels of neurofibromatosis-1 protein and ras oncogene mutations in patients with acute myeloid leukemia and myelodysplastic syndrome. Cancer 97(2):441-9.

[17] Metheny, L.J., Cappione, A.J. and Skuse, G.R. (1995) Genetic and epigenetic mechanisms in the pathogenesis of neurofibromatosis type I. J. Neuropathol. Exp. Neurol. 54, 753-60.

[18] Niles, I. Mapping WordNet to the SUMO Ontology. Teknowledge Technical Report, 2003.

[19] Niles, I. and Pease, A.(2001) Towards a standard upper ontology. Proc. Formal Ontology in Info. Sys.

[20] North, K. (2000) Neurofibromatosis type 1. Am. J. Med. Genet. 97, 119-127.

[21] Oliver, D.E., Hewett, M., Rubin, D.L., Stuart, J.M., Klein, T.E. and Altman, R.B. (2001) Management of data, knowledge, and metadata on the semantic web: Experience with a pharmacogenetics knowledge base. Stanford Med. Informatics Tech. Report .

[22] Oliver, D., Rubin, D.L., Stuart, J.M., Hewett, M., Klein, T. and Altman, R.B. (2002) Ontology development for a pharmacogenetics knowledge base. Pacific Symposium on Biocomputing 7, 65-76.

[23] Okazaki, M., Yoshimura, K., Suzuki, Y., Uchida, G., Kitano, Y., Harii, K., Imokawa, G. (2003) The mechanism of epidermal hyperpigmentation in cafe-au-lait macules of neurofibromatosis type 1 (von Recklinghausen's disease) may be associated with dermal fibroblast-derived stem cell factor and hepatocyte growth factor. Br J Dermatol. 148(4):689-97.

[24] Reish, O., Orlovski, A., Mashevitz, M., Sher, C., Libman, V., Rosenblat, M., Avivi, L. (2003) Modified allelic replication in lymphocytes of patients with neurofibromatosis type 1. Cancer Genet Cytogenet. 143(2):133-9.

[25] Rindflesch, T.C. (1995) Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness. Proc. 5th Annual Dual-use Technol. and Appl. 260-265.

[26] Schwartz, A.S. and Hearst, M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. Pacific Symposium on Biocomputing 8, 451-462.

[27] Shapiro, S.C. and Rapaport, W.J. (1992) The SNePS family. Comp. and Math. with Appl. 23, 243-275.

[28] Sarkar, I.M., Cantor, M.N., Gelman, R., Hartel, F. and Lussier, Y.A. (2003) Linking biomedical language information and knowledge resources: Go and UMLS. Pacific Symposium on Biocomputing 8, 439-450.

[29] Tanabe, L. and Wilbur. W.J. (2002) Tagging gene and protein names in full text articles. Proc. Workshop Nat. Lang. Process. In Biomed. Domain, 9-13.

[30] Thomson, S.A., Fishbein, L. and Wallace, M.R. (2002) NF1 mutations and molecular testing. J. Child Neurol. 17, 555-561.