Natural Language Query in the Biochemistry and Molecular Biology Domains Based on Cognition Search™

Elizabeth J. Goldsmith[†]||, Saurabh Mendiratta[†], Radha Akella[†,] and Kathleen Dahlgren||[§]
[†]Department of Biochemistry, The University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8816. [§] Cognition Technologies, Inc, 6133 Bristol Parkway, Culver City, CA 90230.

Abstract

Motivation: *With the tremendous growth in scientific literature, it is necessary to improve upon the standard pattern matching style of the available search engines. Semantic NLP may be the solution to this problem. Cognition Search (CSIR) is a natural language technology. It is best used by asking a simple question that might be answered in textual data being queried, such as MEDLINE. CSIR has a large English dictionary and semantic database. Cognition's semantic map enables the search process to be based on meaning rather than statistical word pattern matching and, therefore, returns more complete and relevant results. The Cognition Search engine uses downward reasoning and synonymy which also improves recall. It improves precision through phrase parsing and word sense disambiguation.*

Result: *Here we have carried out several projects to "teach" the CSIR lexicon medical, biochemical and molecular biological language and acronyms from curated web-based free sources. Vocabulary from the Alliance for Cell Signaling (AfCS), the Human Genome Nomenclature Consortium (HGNC), the United Medical Language System (UMLS) Meta-thesaurus, and The International Union of Pure and Applied Chemistry (IUPAC) was introduced into the CSIR dictionary and curated. The resulting system was used to interpret MEDLINE abstracts. Meaning-based search of MEDLINE abstracts yields high precision (estimated at >90%), and high recall (estimated at >90%), where synonym information has been encoded. The present implementation can be found at http://MEDLINE.cognition.com.*

Contact*:*
Elizabeth.goldsmith@UTsouthwestern.edu
Kathleen.dahlgren@cognition.com

Introduction

With the increasing complexity of Biomedical literature, several labs and companies have attempted to develop better search engines for MEDLINE (1-5). A few free sources are visible on the web e.g. Google Scholar (http://scholar.google.com/), Highwire press (http://highwire.stanford.edu/lists/freeart.dtl) and Medscape(http://www.medscape.com/home) whereas other relatively commercial sources of this information is present at Scopus (http://www.scopus.com/scopus/home.url), Ovid (http://www.ovid.com/site/index.jsp), and Infotrieve (http://www4.infotrieve.com/newMEDLINE/search.asp). We think that semantic NLP is require to properly access the biomedical literature, a view shared with many others(4, 6-13). To our knowledge, however Cognition semantic NLP is the only technology that has thoroughly unraveled the full complexity of ordinary English. The architecture and databases of the software are such that multiple meanings of ordinary words and synonymy are resolved. The goal in search technology is to create software that finds all the desired information (full recall) without producing undesired information (high precision). Cognition's Semantic MEDLINE has the ability to target and locate specific data that are otherwise hidden in masses of information. Its comprehensive Semantic Map includes words, phrases and idioms. CSIR is also able to select senses of ambiguous words, giving much better results than pattern matching.

Architecture of CSIR™

CSIR™ is a natural language processing (NLP) technology that has been under development for several years. The patented meaning-based architecture and methods have been described previously (14-16). The technology contains a broad semantic map of English based on word senses, their synonyms (6), hypernyms (higher nodes in an ontology) (7) and sense contexts. The CSIR Indexer uses its NLP component to build a cognitive model of the text in which all of the concepts (word meanings) of a document are indexed as well as word strings. The NLP component relies on its dictionary, semantic map, and morphological and syntactic tags (fig.1). At search time, CSIR interprets the query for meaning, and searches for the meaning of the query in the concept index.

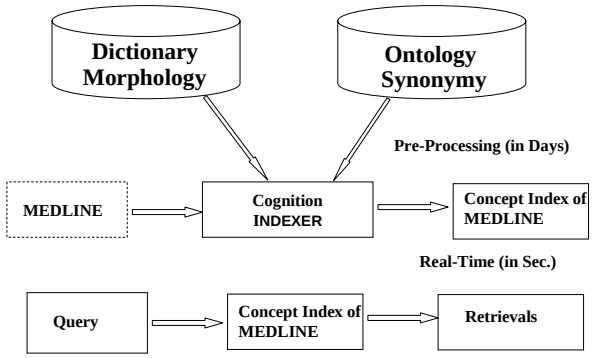

Figure 1: Architecture of CSIR

Since the original descriptions of this technology, significant improvements have been introduced, including sense disambiguation (8), phrase parsing (17), data compression and speed upgrades (18). The morphology and tokenization components were built in-house (patent pending).

The software also uses relatively simple algorithms for phrasal parsing and document relevancy to improve precision. Demonstrations of CSIR are available at http://medline.cognition.com and http://wikipedia.cognition.com. The search engine should be used asking a straightforward question that might be answered in MEDLINE, such as "Oxidative stress in plants," "spectroscopy of amidohydrolases," or "Depression in aging." Retrieval time on the 17 million MEDLINE abstracts is sub-second on Xeon Dual Core 3.0 GHz computers with 1 GB of RAM.

Methods

Ontology:

To augment the ontology for Biochemistry and Molecular Biology, a top ontology was constructed by hand, based upon our own domain knowledge. Websites of curated biomedical terminology were crawled to obtain a complete list of their ontological attachments. These were then mapped to our top ontology by hand.

Lexical and Concept Thesaurus Augmentation:

Biomedical terminological databases were crawled and the vocabulary (terms, phrases and acronyms) extracted, along with their synonyms and ontological classes, where available. All vocabulary was checked for frequency in the MEDLINE abstracts and any items with fewer than 20 occurrences were deleted. Redundancy with the current dictionary was checked automatically, and redundant items curated by hand. Specialized programs were written to crawl each website. Curated terms, synonyms and attachments were automatically added to the CSIR semantic map. Acronym spell-outs were used as sense contexts for acronym meanings (9).

Precision and Relative Recall Test of CSIR vs Pubmed.

We formulated 50 queries for the MEDLINE abstracts. The total number of CSIR retrievals was recorded, and the relevance evaluated for the top 10 and top 20 retrievals, as assessed by the UT Southwestern team. The same queries were posed to PubMed for comparison (in a Boolean format: "genetic" AND "interaction" AND "BCL2"). Relative recall was assessed by taking as full recall the largest number of relevant results found by either search engine. The queries used can be seen on the E.J. Goldsmith Lab webpage (http://hhmi.swmed.edu/Labs/bg/Cognition).

Results

Scale and Scope:

CSIR functions optimally when the semantic map "knows" the vocabulary in the documents. At the initiation of this project, a lexical evaluation of MEDLINE showed that CSIR was missing 66,000 tokens (words). Estimates of the total number of Biomedical terms is over a million, a much larger number, mostly phrases (10). Before this work, the CSIR Lexicon contained about 20,000 medical or biological terms (species, cells, anatomy, etc.). Here we added about 85,000 protein names, 35,000 chemical names, an ontology for Biochemistry and Molecular Biology possessing 2,400 nodes, and over 30,000 biomedical synonym classes. Together with other ongoing lexical augmentations, the detailed description of the entire Cognition semantic map is present in Table 1.

Cognition's Semantic Map (Based on Computational Linguistic Science)	
Word Stems	**506,000 Word stems**
Words and Phrases	**536,000 Word senses or concepts**
Meanings in context	**4,000,000 Semantic contexts**
Different Word Meanings	**17,000 Ambiguous word definitions**
Complex Word Series Meanings	**191,000 Phrases**
Ontology or Taxonomy	**7,000 Nodes & 536,000 Terms**
Synonyms	**76,000 Thesaural concept groups**

Table 1: Cognition Dictionary by numbers

Ontology for Biochemistry and Molecular Biology

Ontologies need to be established at the desired granularity. We defined a top ontology for the Biochemical and Molecular Biology domain that serves as a basis for capturing finer, more desired ontological nodes. Our top ontology, primarily for molecular entities, resembles SEMEDA (7), or TAMBIS (11). The very top of our ontology discriminates 'proteins," laboratory procedures," etc.; an intermediate level of protein and gene names was inspired by the ontology in the AfCS (eg. "binding protein," "g-protein", transcription-factors), and by an ontology of terms in the HGNC that categorizes proteins and genes. (Table 2)

Table 2A: Ontology of Biochemical and Molecular Biology

A. **Piece of the Top Ontology for Biochemistry**
Macromolecule-node
Protein-stuff
antibody
binding protein
enzyme
Nucleic-acid
Laboratory-procedure
electrophoresis
Spectroscopy

B. **Ontology for protein kinases**
protein-kinases
protein-histidine-kinases
serine-threonine-kinases
AGC-kinases
STE-kinase
Tyrosine-kinase
ACK-kinase
EGFR-kinase
Tyrosine-Like-Kinase
MLK-kinase
RAF-kinase

Table 2B: Finer grained Protein Kinases ontology.

Introducing new language from existing databases:

Web-based sources of biomedical terminology were: acronyms from http://medstract.med.tufts.edu (6), the molecules and genes defined by the AfCS database (19), the Human Genome Nomenclature Consortium (20), the UMLS Metathesaurus and the International Union of Pure and Applied Chemistry (IUPAC) enzyme names. The acronym database and UMLS were selected for their wide coverage. We selected the AfCS and HGNC databases because the curators captured natural word usage, and have encoded a gross molecular ontology as well as some synonymy. The IUPAC database was chosen because the ontology has been constructed carefully. Some of the larger databases were avoided because we noted numerous errors and short and redundant acronyms, requiring too much curation. Many biomedical acronyms are ambiguous. Further, since some acronyms were added to the semantic map in earlier projects, a challenge was to add only new senses (21). We chose to use the database published at http://medstract.med.tufts.edu. We curated 16,256 acronyms, removing rarely used acronyms (usage cutoff of 20), and very redundant acronyms. This resulted in 15,657 acronyms with 16,858 total meanings.

We introduced vocabulary from the UMLS Metathesaurus. We built a map from the Metathesaurus ontology to our existing ontology, and then introduced the UMLS vocabulary into the lexicon automatically. Multi-sense words were inspected by a linguist to prevent duplication. Synonyms, with the appropriate senses, were introduced to the Concept Thesaurus automatically.

Normalization included removal of plurals, redundant capitalized versions, and re-ordered versions. Automatic discovery of additional normalization rules, as in Wellner (2005) and Yoshimasa (2008)(22, 23) would be a further step. This database includes both nouns and verbs covering biological sciences and medicine, amounting to 88,423 word senses, and 76,816 synonyms.

We then obtained additional word senses, all nouns, from the Alliance for Cell Signaling (www.alliance.org) (19). This source is current, curated and offers ontological entries, giving 15,661 new or improved word senses. The adoption of this vocabulary was accomplished through a combination of automated tasks and expert curation. Duplicates were curated. Unknown vocabulary was then added to the semantic map automatically, including ontological attachments and synonyms. Data from the HGNC (www.genenames.org) (20) has also been partially introduced. About 30 ontologies of protein families in HGNC have been imported, including AKAPs, ADAM proteases, bcl, BRCA, channel proteins, P450s, tubulins, ubiquitin ligases, phosphatases, TNF-receptors, histones, SMADs, and so on. We also introduced the IUPAC enzyme names and EC numbers, over 6,000 names. These were chosen because of the well-thought-out ontology that may be accessed with the EC numbers. A difficulty with this augmentation is the lack of natural language usage and lack of synonymy. In a separate project we introduced natural language terms by finding synonyms for the EC numbers in the UMLS.

Vocabulary growth

At the beginning of this project, there were 66,000 missing tokens (words). At present, we have completed the addition and curation of all words with a frequency greater than 35 (fig. 2), and there are now 5,000 with frequency greater than 20 to add.
MEDLINE abstracts were also searched to find verbs, which were curated to find words (such as express, silence, translocate, spin, sandwich, bait, prey) that have domain specific-meanings. This project has led to 225 new word senses. The added verb definitions contribute to improved precision, and will be useful when full sentence parsing is included in CSIR (12).

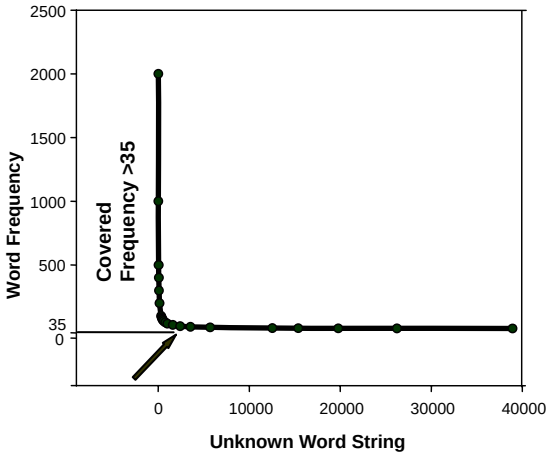

Figure 2 shows the Coverage of MEDLINE

Precision and Recall Test:

50 typical queries for MEDLINE were formulated as simple questions in the areas of biochemistry, molecular biology and medicine. The UT Southwestern team tabulated the relevance of the retrievals in http://MEDLINE.cognition.com. The reader is perhaps the best judge of the performance of the search engine. However, we compared Cognition's retrievals with those of Pubmed (http://pubmed.com). To make the evaluation manageable, we used the "relative recall" technique, wherein full recall is estimated as the greatest number of retrievals achieved by either search engine. For example, one of the queries was "genetic correlates of alcoholism". Of the first twenty CSIR retrievals, 16 were relevant. Thus CSIR's precision was 16/20 or .8. The total number of retrievals for CSIR was 1,436. To extrapolate the good retrievals, we multiplied the precision ratio .8 times 1,436 to yield extrapolated recall of 1,149. A similar calculation for Pubmed was .3 precision, a total of 44 retrievals, to yield extrapolated recall of 13 (Table 3).

Of the two extrapolated recall numbers, CSIR's is greater by inspection, so it is taken to be full recall on this query. Then recall for the two search engines on this query is calculated: CognitionSearch 1,149/1,149 or 1, Pubmed 13/1,149 or .01. Precision and recall ratios for all 50 queries are averaged to calculate the overall precision and recall.

Table 3 Precision and Recall: Comparison between Cognition and Pubmed.

Cognition vs MEDLINE search	Cognition good/20	Cognition bad/20	Total	Pubmed good/20	Pubmed bad/20	Total
Genetic correlates of alcoholism	18	2	1436	6	14	44
DNA repair and aging	17	3	1220	11	9	1265
Drugs for fibromyalgia	17	3	1484	9	11	220
Genetic interactions of BCL2	18	2	876	8	11	19
Oxidative stress in plants	18	2	3122	9	11	3197
spectroscopy of amidohydrolases	17	3	861	7	13	1142
Benzene induced neuropathy	18	2	220	6	1	7
Birth defects from glycol ether	16	4	20	13	7	61
Depression in aging	19	1	13381	7	13	3658
Symptoms of type II diabetes mellitus	18	2	241	7	13	24704
Menopause and depression	18	2	696	11	9	1146
Treatment for bronchiectasis	18	2	2163	6	14	3207
OCD and anorexia	20	0	176	14	6	247
Proteolysis in SARS virus entry	4	0	4	2	0	2
Total	280	60	18433	125	127	34080
	Cognition			MEDLINE		
Precision	0.90			0.50		
Recall (*Assume total recall is the total of the cognition retrievals)	0.99			0.54		

Bootstrapping ontological attachments:

Most of the vocabulary derived from the acronym database and the UMLS had poor (very general) ontological attachments (eg, “amino-acid”). About 80,000 of 136,000 protein names were poorly attached. Attachments of well-classified words were spread to their synonyms resulting in 20,000 better attachments. A bootstrapping method took substrings as triggers; for example, “helix-loop-helix” as a substring of “transcription-factor-15-basic-helix-loop-helix” suggests an attachment to the node “helix-loop-helix.” This attachment was then assigned to the synonyms “bHLH-EC2-protein” and “paraxis”.

Discussion

We think that the natural language approach of CSIR has an important role in future access to textual information in the Biomedical domain. This effort is our first pass at introducing Biochemical and Molecular Biology terms into the CSIR lexicon. Other sources of new words will come from tracking user queries, evaluation of MEDLINE, and other curated databases. Efforts directed toward database integration may provide useful definitions, synonymy and ontology in molecular biology (13). We also plan to introduce additional parsing functions (24), (12) which should improve the precision of Cognition Search. CSIR works equally well on full-text as on abstracts. This work contributes to precise interpretation of biomedical texts for purposes of search (1, 3, 25), research (4) and data mining (2, 26).

Uses and Applications of CSIR:

It is useful to review which linguistic processes produce these improved results. Morphology improves recall, so that the user can state a query term in one of its morphological variants, and CSIR automatically finds all other forms, as in phosphorylate and phosphorylation. Synonymy improves recall because one member of a synonym class retrieves documents with any of its members, as in “CD116,” "GMHCFS receptor alpha subunit," etc. Ontological reasoning improves recall as the software reasons down from higher-level concepts to lower-level concepts. For example, you can query "what MAP kinase phosphorylates ATF2" and get documents with “ERK” and “p38” which are kinds of MAP kinases. Sense disambiguation improves precision because only the documents that contain the query terms in the meanings intended by the user are retrieved. Phrase parsing improves both precision and recall. It improves precision by avoiding retrievals that happen to contain parts of a phrase in various positions, but not as the phrase. So "RNA", "binding" and "protein" might all appear in an abstract that has nothing to do with RNA binding proteins. It improves recall because it enables the mapping of synonym relations between phrases, and between phrases and acronyms, as in "TUBB" and "beta-tubulin".
Biomedical language also possesses ontological relationships for proteins, genes, the Tree-of-Life animals, diseases, etc. CSIR includes the function of downward reasoning in ontologies. Thus, CSIR NLP technology can help to solve problems in medicine by finding material about specific instances of general concepts such as “heart disease medicine”.

Areas for improvement

Precision is lowered when words are difficult to disambiguate, such as "Bad", which is an apoptosis protein, but at present is recognized as the ordinary English "bad". It will be relatively easy to address missing terms since we know there are still 5000 individual terms used in MEDLINE with a frequency of 20 or more that we need to define. We will use the methods of Tsuruoka (27) for future term

recognition, synonymy expansion and evaluation of coverage.

Acknowledgements

We thank Ron Taussig for pointing out the Alliance for Cell Signaling website and other discussions. UMLS resources licensed (number 21817A334). The work in E. J. Goldsmith's group was carried out under contract with Cognition, Technologies, Inc.

References:

1. Vanhecke TE, Barnes, M.A., Zimmerman, J., Shoichet, S. PubMed vs. HighWire Press: A head-to-head comparison of tow medical literature search engines. Computers in Biology and Medicine. 2007; 37:1252-8.
2. Divoli A, Attwood, T.K. "BiolE sentences - Extracting informative sentences from the biomedical literature." Bioinformatics 2005; 21(9):2138-9.
3. Doms A, Schroeder, M. "GoPubMed: exploring PubMed with the Gene Ontology". . Nucleic Acids Research 2005; 33.
4. Fontelo P, Liu, F., Ackerman, M. "askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. BMC Medical Informatics and Decision-Making 5:5, 2005.
5. Matthew E. Falagas, 1, Eleni I. Pitsouni, George A. Malietzis and Georgios Pappas. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB 2008; 22:338-42.
6. Wren JD, Chang JT, Pustejovsky J, Adar E, Garner HR, Altman RB. Biomedical term mapping databases. Nucleic Acids Res. 2005 Jan 1; 33(Database issue):D289-93.
7. Kohler J, Schulze-Kremer S. The semantic metadatabase (SEMEDA): ontology based integration of federated molecular biological data sources. In Silico Biol. 2002; 2(3):219-31.
8. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics. 2001; 17 Suppl 1:S97-106.
9. Yu H, Kim, W., Hatzivassiloglou, V., and Wilbur, W. Disambiguating biomedical abbrevations. ACM Transactions on Information Systems (TOIS). 2006; 24(3):380-404.
10. Bodenreider O. Lexical, terminological and ontological resources for biological text mining. Ananiadou S, McNaught, J., editor: Artech House; 2006.
11. Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A. An ontology for bioinformatics applications. Bioinformatics. 1999 Jun; 15(6):510-20.
12. Pustejovsky J, Castano J, Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput. 2002b:362-73.
13. Philippi S, Kohler J. Using XML technology for the ontology-based semantic integration of life science databases. IEEE Trans Inf Technol Biomed. 2004 Jun; 8(2):154-60.
14. Dahlgren K, McDowell, J., and Stabler, E.P. Knowledge Representation for Commonsense Reasoning with Text. Computational Linguistics. 1989; 15:149-70.
15. Dahlgren K. Interpretation of Textual Queries Using a Cognitive Model. Ehrlbaum; 1992.
16. Dahlgren K, editor. Improving Precision and Recall with Linguistic Semantics. Proc Semantic Technology Conference; 2007; San Jose, CA.
17. Kornai A. Mathematical Linguistics. Springer; 2008.
18. Witten IH, Moffat, A.M., and Bell, T.C. Managing Gigabytes of Data. New York, NY. Morgan Kaufmann.; 1999.
19. Gilman AG. Cross talk: interview with Al Gilman. Mol Interv. 2001 Apr; 1(1):14-21.
20. Wain HM, Lush M, Ducluzeau F, Povey S. Genew: the human gene nomenclature database. Nucleic Acids Res. 2002 Jan 1; 30(1):169-71.
21. Wren JD, Garner HR. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods Inf Med. 2002; 41(5):426-34.
22. Wellner BC, J and Pustejovsky, J. . "Adaptive string similarity metrics for biomedical reference resolution". Proc ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics; 2005. p. 9-16
23. Yoshimasa T, McNaught, J and Ananiadou, S. . "Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinformatics 9(Suppl 3). 2008(S2).
24. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics. 2001; 17 Suppl 1:S74-82.
25. Eaton AD. HubMed: a web-based biomedical literature search interface. Nucleic Acids Research 2006; 34.
26. Lee S, Yang, L., Jianrong, L., Friedman, C., Lussier, Y.A. . "Discovery of protein interaction networks shared by diseases". . Pacific Symposium on Biocomputing; 2007. p. 76-87.
27. Tsuruoka Y, McNaught J, Tsujii J, Ananiadou S. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics. 2007 Oct 15; 23(20):2768-74.