

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

Elizabeth J. Goldsmith^{†||}, Radha Akella[†], Saurabh Mendiratta[†], and Kathleen Dahlgren^{||§}
[†]Department of Biochemistry, and Pharmacology, The University of Texas Southwestern
Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8816. [§]
Cognition Technologies, Inc, 6133 Bristol Parkway, Suite 350, Culver City, CA 90230.

Abstract

CognitionSearch™ (CSIR™), developed by Cognition Technologies, Inc., is a natural language processing (NLP) search engine with a broad semantic map of English. Web-based sources of language and acronyms in Biochemistry and Molecular Biology were incorporated semi-automatically into the CSIR™ lexicon. 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. Multiple word senses were kept distinct. Synonyms and phrases were recorded. Together with prior encoding of medical, tree-of-life, and anatomical words, CSIR covers a spectrum of "Biomedical" disciplines. 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.

Introduction

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 (1-3). The technology contains a broad semantic map of English based on word senses, their synonyms (4), hypernyms (higher nodes in an ontology) (5) 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. At search time, CSIR interprets the query for meaning, and searches for the meaning of the query in the concept index.

Since the original descriptions of this technology, significant improvements have been introduced, including sense disambiguation (6), phrase parsing (7), data compression and speed upgrades (8). The morphology and tokenization components were built in-house (patent pending). The software also uses relatively simple algorithms for concept clustering and document relevancy to improve precision. Demonstrations of CSIR are available at <http://WIKIPEDIA.cognition.com>.

Search of Biomedical texts stands to be improved using CSIR. CSIR has the architecture to handle multiple synonyms for heterogeneously named molecules and genes. Biomedical language also possess ontological relationship, for proteins, genes, Tree-of-Life, diseases, etc, and again, CSIR includes the function of downward reasoning in ontologies. Here we have carried out several projects to "teach" CSIR Biochemistry and Molecular Biology terms from curated web-based free sources, and indexed MEDLINE. The present implementation can be found at <http://MEDLINE.cognition.com>. Retrieval time on the 17 million MEDLINE abstracts is sub-second on Xeon Dual Core 3.0 GHz computers with 1 GB of RAM. The search engine should be used asking a straightforward question that might be answered in MEDLINE, such as "genetic correlates of alcoholism," "ligands for hippocampus GPCRs," or "spectroscopy of PLP enzymes." The goal in search technology is to create software that finds all the desired information (full recall) without producing undesired information (high precision). Typically, either recall is high and precision low, or the reverse. Using the natural language techniques described here, both can be high.

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. 50 queries for the MEDLINE abstracts were formulated by biochemists. 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 “correlate” AND “alcoholism”). 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 E.J. Goldsmith Lab’s 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, and ontology for Biochemistry and Molecular Biology possessing 2,400 nodes, and over 30,000 biomedical synonym classes. Together with other lexical augmentations ongoing at Cognition, Inc., the entire semantic map currently has 506,000 word stems, 536,000 senses, 75,000 synonym classes and 7,564 ontological nodes in all language domains.

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 (5), or TAMBIS (11). The very top of our ontology discriminates ‘proteins," laboratory procedures," etc.; an intermediate level protein and gene names was inspired by the ontology in the AfCS (eg. "binding protein," "g-protein", transcription-factors), and an ontology of terms in the HGNC that categorizes proteins and genes. (Table 1A)

Table 1. Ontology of Biochemical and Molecular Biology

A. Piece of the Top Ontology for Biochemistry
Macromolecule-node
Carbohydrate-node
Protein-stuff
antibody
binding protein
enzyme
receptor
transcription factor
Inhibitor protein
Signaling protein
Ubiquitin pathway
Nucleic-acid
Laboratory-procedure
electrophoresis
mass-spectrometry
spectroscopy
B. Piece of Finer-grained Ontology for protein kinases
protein-kinases
protein-histidine-kinases
serine-threonine-kinases
AGC-kinases
STE-kinase
CMGC-kinase
CK2-kinase
CAMK-kinase
Tyrosine-kinase
ABL-kinase
ACK-kinase
EGFR-kinase
Tyrosine-Like-Kinase
MLK-kinase
RAF-kinase
TGFBR-kinase

Table 1B gives a piece of finer grained ontology.

Introducing new language from existing databases. Web-based sources of biomedical terminology were: acronyms from medstract.med.tufts.edu (4), the molecules and genes defined by the AfCS database (12), the Human Genome Nomenclature Consortium (13), 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 (14). 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 then introduced vocabulary from the UMLS Metathesaurus. We built a map from the Metathesaurus ontology to the 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. This database includes both nouns and verbs covering biological sciences and medicine, amounting to 88,423 word senses, and 76,816 synonyms. The ontological attachments in the UMLS are very general (discussed below).

We then obtained additional word senses, all nouns, from the Alliance for Cell Signaling (www.alliance.org) (12). 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) (13) has also been partially introduced. Of the 150 or so ontologies of protein families in HGNC, about 30 have been imported, including AKAPs, ADAM proteases, bel, BRCA, channel proteins, P450s, tubulins, ubiquitin ligases, phosphatases, TNF-receptors, histones, SMADs, and so on. We also introduced the IUPAC enzyme names, over 6,000 names. These were introduced for the nice 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.

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” suggests node “helix-loop-helix.” This attachment was then assigned to the synonym bHLH. The remaining 60,000 problem terms will be the focus of future work.

Missing word and verb approaches: At the beginning of this project, there were 66,000 missing tokens (words). However, analysis of MEDLINE showed that the usage frequency was surprisingly low for the vast majority of the tokens; 60,000 are used less than 10 times. Only 552 missing words were present more than 100 times and 2500 used more than 20 times in MEDLINE. At present, we have completed the curation of 1200 of the most frequent missing words, and plan to introduce the remaining words with a frequency greater than 20.

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 leads 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 (15).

Precision and Recall Test: 50 queries for MEDLINE that the UT Southwestern team believed would be typical of molecular biology searchers were

formulated as simple questions in the areas of biochemistry and medicine. The UT Southwestern team tabulated the relevance of the retrievals in <http://MEDLINE.cognition.com>. Both precision and recall of the retrieval for these queries was over 90%. The reader is perhaps the best judge of the performance of the 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. By inspection, of the two extrapolated recall numbers, CSIR's is greater, 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.

Uses of CSIR: The aspects of the linguistic processing that produce these improved results are reviewed here. 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 caught-catch, or phosphorylate-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 recognition 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, as in "protein tyrosine kinase 9 related protein" and "protein tyrosine kinase 9-like", and between phrases and acronyms, as in "TUBB" and "beta-tubulin".

Areas for improvement.

Retrieval failures derive from several sources. Recall is worsened when a word is not in the dictionary. CSIR can find the string, but cannot perform linguistic reasoning. 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". Also, some synonym classes are broader in ordinary English than in the biomedical domain, and need to be narrowed for Biomedical search.

It will be relatively easy to address missing terms since we know there are still 1300 individual terms used in MEDLINE with a frequency of 20 or more that we need to define. Assuming that we have most phrases and synonyms is more difficult. We will use the methods of Tsuruoka (16) for future term recognition, synonymy expansion and evaluation of coverage.

Discussion and future work.

We think that the natural language approach of CSIR has an important role in future access to textual information in the Biomedical domain. The effort described here is our first pass at introducing Biochemical and Molecular Biology terms into the CSIR lexicon. We will continue to incorporate Biochemical terms from the best-curated web sites. Other sources of inspiration for new words for the dictionary 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 (17). We also plan to introduce additional parsing functions (18), (15) which should improve the precision of CognitionSearch. CSIR works equally well on full-text. In addition to the current client-server technology, the software will be made available for desktop search.

Acknowledgements:

We thank Ron Taussig for pointing out the Alliance for Cell Signaling Website and other discussions. UMLS resources licensed (number 21817A3334). The work in E. J. Goldsmith's group was carried out under contract with Cognition Technologies, Inc.

References:

1. Dahlgren K, McDowell, J., and Stabler, E.P. Knowledge Representation for Commonsense Reasoning with Text. *Computational Linguistics*. 1989;15:149-70.
2. Dahlgren K. *Interpretation of Textual Queries Using a Cognitive Model*.: Ehrlbaum; 1992.
3. Dahlgren K, editor. *Improving Precision and Recall with Linguistic Semantics*. Proc Semantic Technology Conference; 2007; San Jose, CA.
4. 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.
5. 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.
6. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*. 2001;17 Suppl 1:S97-106.
7. Kornai A. *Mathematical Linguistics*.: Springer.; 2008.
8. Witten IH, Moffat, A.M., and Bell, T.C. *Managing Gigabytes of Data*. New York, NY.: Morgan Kaufmann.; 1999.
9. Yu H, Kim, W., Hatzivassiloglou, V., and Wilbur, W. Disambiguating biomedical abbreviations. *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. Gilman AG. Cross talk: interview with Al Gilman. *Mol Interv*. 2001 Apr;1(1):14-21.
13. Wain HM, Lush M, Ducluzeau F, Povey S. Genew: the human gene nomenclature database. *Nucleic Acids Res*. 2002 Jan 1;30(1):169-71.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.