Recognizing Gene and Protein Function in MEDLINE Abstracts

Richard D. Hull, Larry F. Waldman[†]
Axontologic, Inc.
12565 Research Parkway, Suite 300
Orlando, FL 32826
hull@axontologic.com

Abstract

Identification of genes and proteins that affect biological function in humans and other organisms is a critical step in the discovery of new medicinal therapies. Automatic recognition of MEDLINE abstracts that describe gene/protein function would be of tremendous benefit to researchers in industry, government, and academia. Our approach uses simple syntax and domain semantics to both identify sentences from MEDLINE abstracts that suggest gene function and to rank those abstracts by a measure of how many appropriate function instances they contain.

Introduction

Identification of genes and proteins that affect biological function in humans and other organisms is a critical step in the discovery of new medicinal therapies. Automatic recognition of MEDLINE abstracts that describe gene/protein function would be of tremendous benefit to researchers in industry, government, and academia. For example, drug discovery projects often begin with the identification of one or more disease-associated protein targets. Pharmaceutical biologists spend a large portion of their time combing the literature for research articles discussing novel protein targets. Automating this element of their jobs has the potential to result in accelerating the discovery process and reducing costs.

[†]Larry Waldman, School of Computer Science, Carnegie Mellon University, larrywaldman@cmu.edu.

The first Genomics Track of the 12th Text Retrieval Conference (TREC-12) was designed to provide a forum for those interested in developing systems capable of addressing the challenges posed by automatic recognition of gene and protein function. Axontologic and twenty-four other organizations from academia, government and industry participated in the primary task of the Track during the summer and fall of 2003. The official results of all of the participants are available on the TREC website.[1]

This paper begins with an overview of the primary task of the Genomics Tracks. It describes our natural language processing inspired approach and discusses the results of our system. Finally, our conclusions are presented.

Primary Task

The primary task of the TREC-12 Genomics Track began with the release of the Track training set comprised of the task documents, 50 training topics and the training relevancy judgments for those 50 topics. These training resources were made available to the participants in May of 2003. Participants were encouraged to use the training set to understand the nature of the task and develop their systems.

The task documents or corpus contained over 525,000 MEDLINE abstracts indexed between April 2002 and April 2003. The 50 training topics (or queries) were the gene names from 50 LocusLink records. The National

Library of Medicine's LocusLink database "provides a searchable interface to curated sequence and descriptive information about genetic loci."[2] One component of a LocusLink record is its GeneRIF (Gene References into Function), a list of concise statements of the gene's function with associated MEDLINE references. The relevancy judgments for the 50 training topics were those MEDLINE abstracts referenced in the corresponding LocusLink GeneRIFs.

Fifty test topics were released in late June. Each topic contained the gene names from a LocusLink record that was not part of the training set. Participants had until August 4th to submit up to two test runs comprised of a ranked list of at most 1000 MEDLINE abstract identifiers for each of the 50 topics. The results of these runs were judged by the TREC evaluators and returned to the participants two weeks later.

Methodology

Our goal was to explore the use of simple syntax and domain semantics for recognizing gene/protein function and to lay the foundation for competition in future TREC events. Our system is automatic and combines domain independent processing elements with a domain-specific lexicon.

After review of the training set it became clear that there were three challenges to be addressed. First, the gene names provided in the training topics were not comprehensive, i.e., some of the abstracts in the answer set did not mention the given names. Second, using gene names resulted in many false positives, because many of the retrieved documents did not discuss gene function. Third, LocusLink records are organism-specific, therefore, the system had to filter out abstracts that discuss gene function in other organisms. A fourth issue dealing with false negatives could not be addressed directly within the context of the task.

We used the MG (Managing Gigabytes) system[3] to index the MEDLINE corpus. A strategy for heuristic query expansion was developed using a generative grammar. Query terms were expanded using simple grammar rules to handle hyphenation, complex punctuation and common gene name variations, e.g., "alpha-1a adrenergic receptor" vs. "adrenergic receptor alpha-1a." The expanded set of gene names was then fed to MG to retrieve a list of candidate abstracts from the corpus. We limited the number of retrieved documents to the first 5000 in our test runs. The ranking produced by MG was saved for later use.

A proprietary lexicon of terms indicative of gene or protein function was developed by analyzing the training data and other public-domain sources of functional information for genes including the Gene Ontology. The lexicon includes verbs (cleave, inhibit, etc.), nominalizations (activation, regulation, etc.) and adverbs.

The ranked abstracts were parsed into sentences and the sentences were examined to locate query terms and function terms. The abstracts were scored using a function of the frequency of query term/function term pairs found in an abstract's sentences and their proximity to each other, i.e., a gene term adjacent to a function term ("caspase-3 cleaves…") is scored higher than a gene term/function term pair separated by many intervening words. The system differentiates between cases where the gene term is in the subject or object position of the function term and it handles passive constructions.

An additional check was made to verify that the organism of the query was contained in the MeSH terms of each returned abstract. Abstracts were ranked by their functional scores. If no query term/functional term pairs were found in an abstract, then the abstract was ranked using the original MG ranking, after all those abstracts containing query/function pairs.

Results

Axontologic submitted two runs labeled *axon1* and *axon2* – the second run had a slightly more liberal query expansion grammar. The official results are shown in Tables 1 and 2. The results for the two runs were very close with axon2 doing slightly better in average precision (0.3173 compared to *axon1*'s 0.3118).

Table 1. *Axon1* Results.

Recall Level Precision Averages	
Recall	Precision
0.00	0.6372
0.10	0.5764
0.20	0.4725
0.30	0.3930
0.40	0.3420
0.50	0.3020
0.60	0.2504
0.70	0.2268
0.80	0.2070
0.90	0.1699
1.00	0.1468
Average precision over all relevant docs	
non-interpolated	0.3118

Document Level Averages	
	Precision
At 5 docs	0.3200
At 10 docs	0.2400
At 15 docs	0.2120
At 20 docs	0.1890
At 30 docs	0.1493
At 100 docs	0.0742
At 200 docs	0.0437
At 500 docs	0.0203
At 1000 docs	0.0104
R-Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.2935

Table 2. *Axon2* Results.

Recall Level Precision Averages	
Recall	Precision
0.00	0.6361
0.10	0.5764
0.20	0.4697
0.30	0.4056
0.40	0.3543
0.50	0.3141
0.60	0.2598
0.70	0.2394
0.80	0.2100
0.90	0.1748
1.00	0.1518
Average precision over all relevant docs	
non-interpolated	0.3173

Document Level Averages	
	Precision
At 5 docs	0.3200
At 10 docs	0.2500
At 15 docs	0.2187
At 20 docs	0.1930
At 30 docs	0.1520
At 100 docs	0.0748
At 200 docs	0.0433
At 500 docs	0.0201
At 1000 docs	0.0105
R-Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.2959

Discussion

While these results were competitive, there is much more to be done. There were seven queries that axon2 did not meet the median average precision. Six of the seven queries involved problems with the names of the genes given in the topic. For example, in topic 4, the official gene name is given as "guanine nucleotide binding protein (G protein), alpha activating activity polypeptide, olfactory type". There were two MEDLINE abstracts in the answer set for this query, 11901355 and 12037684. The gene names used in these two abstracts are synonyms for the official gene name that were not included in the topic: G-protein alpha(olf), AGalpha(olf), G-protein Golf, and Golf.

Topic 38 held a similar problem for our system. The topic used name "MIP-1-alpha", but we missed answer abstracts that used lexical variations MIP-1 alpha, MIP-1alpha, and (MIP)-1alpha. Our query expansion grammar did not correctly handle cases with multiple dashes.

Clearly the problem of gene name variability is a fundamental issue, much as word choice is in traditional information retrieval. If the MG system did not retrieve an answer abstract, then there was no way for later processing components to compensate for that. We plan to improve our query expansion grammar and to add additional gene synonyms from public domain sources.

On the positive side, there were five topics that one or both of *axon1* and *axon2* returned the best average precision score. Of these five, *axon1* returned an average precision score that was several times greater than the median in topics 10, 14 and 42. To understand what happened with these query topics, we have compared the original MG ranking to the final *axon1*

ranking for the answer abstracts (see Tables 3, 4 and 5). In nearly every case, the rankings of target abstracts were improved by the use of function terms and in many cases the improvements were dramatic.

Table 3. MG and *Axon1* Rankings for Topic 10.

PMID	MG	Axon1
11750880	23	21
11756417	36	1
11913997	108	11
11961237	92	9
12135673	21	3
12187073	252	57
12359731	41	18
12468916	424	69

For example, four target abstracts for Topic 14, PMID's 11865975, 12167626, 12234259 and 12354983, were improved into the top 10 from their original rankings of 687, 1419, 1739, and 3217 respectively. MG returned the maximum 5000 articles for this topic.

Table 4. MG and *Axon1* Rankings for Topic 14.

PMID	MG	Axon1
11865975	687	6
11920569	1082	74
11983915	2169	18
12086670	219	11
12167626	1419	1
12189556	4071	64
12218115	806	20
12234259	1739	9
12270125	-	-
12370314	889	38
12374983	3217	4
12393617	-	-

The strength of our method is in its ability to promote abstracts that discuss the function of the gene over those abstracts that simply mention the gene. In those cases where there are a large number of abstracts containing the topic gene names, our method has the potential to make significant improvements.

Table 5. MG and *Axon1* Rankings for Topic 42.

PMID	MG	Axon1
11756426	743	1
11809755	683	121
11823458	24	108
11877420	891	7
11912192	502	192
11912196	1892	42
11948811	1549	61
11972038	27	37
12068009	90	31
12087104	677	123
12093166	266	126
12101040	222	39
12230982	643	209
12239221	543	198
12431992	488	105
12493631	761	224
12515826	259	9

Conclusions

The Genomics track of TREC-12 provided an environment for developing, testing and evaluating computational methods for ranking abstracts by how well those abstracts describe the function of genes. Participating in TREC has increased our understanding of the problems researchers face and our solutions to those problems will be incorporated in Axontologic's future products.

We intend to bolster the query expansion component of the system to address the problem of not recognizing gene name synonyms in answer abstracts. We are currently investigating more sophisticated methods of scoring that utilize finite state transducers and shallow syntactic

parsing models. It is our belief that even simple methods, however, can be of value to the biomedical community given the large numbers of false negatives found by our system and that of the other competitors.

References

[1] Text REtrieval Conference (TREC) Homepage. http://trec.nist.gov/.

[2] Ward, J. Gene Indexing. NLM Tech Bull. 2002 Sep-Oct;(328):e6.

[3] Managing Gigabytes: Compressing and Indexing Documents and Images, Witten, I. H., Moffat, A, and Bell, T. C., Morgan Kaufmann Publishing, San Francisco, 1999.