Phrases, boosting, and query expansion using external knowledge resources for genomic information retrieval

William Hersh, Ravi Teja Bhupatiraju, Susan Price
Oregon Health & Science University
{hersh,bhupatir,prices}@ohsu.edu

In our TREC Genomics Track work, we focused on domain-specific techniques in attempting to improve retrieval performance beyond a word searching baseline. One set of experiments looked at using phrases based on gene name synonyms with boosting of the canonical name of the gene. Another set assessed query expansion using external knowledge resources.

Query expansion has been a staple of the TREC ad hoc task dating back almost to the inception of TREC, showing consistent benefit when added to a wide variety of baseline techniques, e.g., [1, 2]. In the biomedical domain, however, results have been mixed. While Srinivasan obtained improved retrieval using retrieval feedback (automatic relevance feedback) in a small test collection [3], Hersh et. al. did not find improved retrieval when queries were expanded using thesaurus relationships in the Unified Medical Language System (UMLS) Metathesaurus [4]. Query expansion may be feasible in the genomics domain due to the considerable effort being devoted to creating useful cross-linkages across data sources. The most prominent example is the collection of databases maintained by the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov), a division of the National Library of Medicine (NLM, www.nlm.nih.gov) [5].

Phrases and Boosting

Our first experiments derived from a goal of defining baseline performance for the track training data. We built an IR system around the Lucene search engine with a pre-processor implemented in Python and a batch search facility implemented in Java. Shell scripts tied together these components in a way to allow experiments. The pre-processor converted the formal track queries file to a text file that contained the query terms in a Lucene format with one line per query. The search facility took this file and batched the queries into Lucene. The results were written into a text file in the trec_eval format. The report script called trec_eval with the results and the relevance file (qrels), writing the output to a report text file.

Methods

The simplest baseline approach involved taking all of the names for a gene name and turning them into a single query string of a "bag of words." Table 1 shows the gene names provided from LocusLink for topic 1 of the test data. This was transformed into the following query string:

> cyclin-dependent kinase inhibitor 1A p21 Cip1 CDKN1A P21 CIP1 SDI1 WAF1 CAP20 CDKN1 MDA-6 cyclin-dependent kinase inhibitor 1A cyclin-dependent kinase inhibitor 1A cyclin-dependent kinase inhibitor 1A DNA synthesis inhibitor CDK-interaction protein 1 wild-type p53-activated fragment 1 melanoma differentiation associated protein 6

Table 1 - Gene names for test topic 1 from LocusLink.

1	1026	Homo sapiens	OFFICIAL_GENE_NAME	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
1	1026	Homo sapiens	OFFICIAL_SYMBOL	CDKN1A
1	1026	Homo sapiens	ALIAS_SYMBOL	P21
1	1026	Homo sapiens	ALIAS_SYMBOL	CIP1
1	1026	Homo sapiens	ALIAS_SYMBOL	SDI1
1	1026	Homo sapiens	ALIAS_SYMBOL	WAF1
1	1026	Homo sapiens	ALIAS_SYMBOL	CAP20
1	1026	Homo sapiens	ALIAS_SYMBOL	CDKN1
1	1026	Homo sapiens	ALIAS_SYMBOL	MDA-6
1	1026	Homo sapiens	PREFERRED_PRODUCT	cyclin-dependent kinase inhibitor 1A
1	1026	Homo sapiens	PRODUCT	cyclin-dependent kinase inhibitor 1A
1	1026	Homo sapiens	ALIAS_PROT	DNA synthesis inhibitor
1	1026	Homo sapiens	ALIAS_PROT	CDK-interaction protein 1
1	1026	Homo sapiens	ALIAS_PROT	wild-type p53-activated fragment 1
1	1026	Homo sapiens	ALIAS_PROT	melanoma differentiation associated protein 6

Because our results generated relatively low precision at various points of recall, we looked for ways to decrease the “noise” in our queries. One attempt to improve precision involved the use of phrases. Using the feature of Lucene that allows adjacency of words in a query to be designated by enclosing them in quotes (a common feature across Web search engines), we rebuilt the queries as a series of phrases. One approach involved using only the official gene name in a phrase. The search string for this approach for the gene in Table 1 was:

> cyclin-dependent kinase inhibitor 1A (p21, Cip1)

Another approach converted all of the gene names into phrases. We also discovered some additional performance improvement by using another feature of Lucene, term boosting, which allows designated terms to be assigned added weight. We empirically determined that increasing the weight of the official name phrase by 2.9 gave the best performance. The search string for this approach for the gene in Table 1 was:

> "CDKN1A"^2.9 "P21" "CIP1" "SDI1" "WAF1" "CAP20" "CDKN1" "MDA-6" "cyclin-dependent kinase inhibitor 1A" "cyclin-dependent kinase inhibitor 1A" "cyclin-dependent kinase inhibitor 1A" "DNA synthesis inhibitor" "CDK-interaction protein 1" "wild-type p53-activated fragment 1" "melanoma differentiation associated protein 6"

Results

Table 2 shows the results of these queries for the training and test data. For both data sets, the use of the official name in a phrase improved MAP modestly, while phrases of all names more than doubled it. Boosting added a small gain in MAP. One interesting finding, not only for us but also another group making their training data results public (J. Savoy, Institut interfacultaire d'informatique, Université de Neuchâtel), was that MAP decreased across the board for the test topics, although the relative performance of the different approaches was comparable. Although we have not yet analyzed why this happened, we suspect it has to do with the decision to limit the test queries to genes with three or more GeneRIFs.

Table 2 - Baseline, phrases, and boosted results for training and test data.

Topics	Run	Retrieved	Relevant	Retrieved & Relevant	Mean Average Precision
Training	Names to words	46115	335	143	0.1584
	Official name as phrase	2829	335	100	0.1998
	All names as phrases	12585	335	215	0.3256
	Official name phrase boosted	12583	335	215	0.3351
Test	Names to words	48021	566	294	0.0741
	Official name as phrase	6197	566	220	0.1372
	All names as phrases	14830	566	419	0.1725
	Official name phrase boosted	14820	566	419	0.1747

Query Expansion Using External Knowledge Resources

The bioinformatics community has produced a wealth of publicly available databases that contain various kinds of information such as:

- gene sequences
- gene clustering
- protein products
- microarray data
- apparent function
- association with diseases
- gene expression in various tissue types

Much of this information is available on the Web, often in HTML and sometimes in XML formats. The volume of data is often huge, with much of it stored in databases that can be accessed only in response to queries via Web forms whose actions are to pass the query to a database and display the results on a Web page. In addition to the large number of primary data sources, such as those maintained by NCBI, there are sites that aggregate various kinds of data. A good example is Source, which is published by a research group at Stanford University (source.stanford.edu) and compiles data from at least five different public databases [6]. The aggregation of data in Source, from multiple databases in an easily processed standardized output, led to its selection as the initial source of information for augmenting the queries.

Methods

For the query expansion experiments, we used the boosted run described above as our baseline query for expansion. Each type of information was added to the baseline query for each gene in a separate run, yielding 13 runs in addition to the baseline run. Data were extracted from Source using a collection of Perl programs. The first program automatically filled in the query form and downloaded the resulting web pages to a local file. Another program read the file for each Web page and extracted the data.

Thirteen pieces of data were collected whenever possible, but not all information was available for all genes. For the 50 genes in the test set, the number of genes for which data from a given category was available ranged from two, for the descriptions associated with accession numbers for mRNA sequences in the NCBI Reference Sequence (RefSeq) records, to 49, for the UniGene Cluster ID. Unfortunately, five of the genes were from Drosophila melanogaster (fruit fly) and had no information about them available from Source. Partial data for those queries was obtained manually from the LocusLink and

FLYBASE (flybase.bio.indiana.edu) databases. One query, PTEN, had two gene entries for the same symbol and name. Most of the data was the same for both, but differed in the UniGene Cluster ID and the tissue types in which they were expressed. Data from both was included in the expansion for that gene. Another query, Prkca in Rattus norvegicus, did not have any data in Source. Thus the data in our experiments come from 49 genes; 44 that are human, mouse, or rat, for which information is available from Source, and five fruit fly genes for which partial information was obtained manually. Queries for which a data item was not available were left unexpanded but were included in the run for that data category.

Additional Perl programs created a new query file for each of the 13 types of information and parsed the reports produced from each run in order to extract and collate the results. These files were then input into Lucene and the results passed with the qrels file to trec_eval.

Results

Our results are summarized in Table 3. None of the thirteen sources of information improved MAP of documents when added to each of the queries for which it was available. Some information categories actually caused a sizeable decrease in MAP. Direct comparisons of MAP must take into account the number of queries for which a data category was available. If data were available for only a few queries, the effect on MAP across all queries would be limited no matter how much the approach might improve it for individual queries. To mitigate this limitation, for each type of information added to the query, we also calculated how many queries were either improved or made worse by the additional information. Despite the overall decline in MAP, as many as one third of queries did see an improvement with the added information, as shown in Table 4.

Addition of identification and accession numbers had little effect on retrieval statistics. As might be expected, the categories that contained the most text words were the most likely to cause changes in retrieval statistics, both positive and negative. In general, the LocusLink summary contained the most text words, but had an almost uniformly negative effect on retrieval performance. That MAP was not the worst of all categories probably reflects the fact that it was only available for 13 of 50 queries. Addition of the top ten tissue types in which the gene is expressed had the most deleterious effect on MAP, and also had a negative effect on most of the queries for which it was available. The information from the Gene Ontology consisted primarily of words or phrases. Effects were both positive and negative, but the negative results outweighed the positive. The textual information about protein function from the SwissProt database had similarly mixed results.

Discussion

Our experiments established baseline results for the TREC Genomics Track. Designating names as phrases improved performance, especially when we used all names for the genes and boosted the weight of the official name.

Table 3 - The performance for query expansion using external resources. The baseline results consist of the best run from the phrases and boosted approach described above. Each subsequent row represents the results when a single piece of information was added to the query for each gene for which the information was available.

Fields Added to Query	Queries expanded	Retrieved	Relevant	Relevant & Retrieved	Mean Average Precision
Baseline	50	14820	566	419	.1747
Chromosome Location	44	25283	566	409	.1655
Locus Link Summary	13	25386	566	313	.1414
SwissProt Accession No.	41	14820	566	419	.1747
Protein function (from SwissProt database)	40	42092	566	317	.1268
Relationships to disease (from SwissProt database)	7	18967	566	350	.1640
Molecular functions of the gene product (from Gene Ontology)	43	45491	566	334	.1400
Biological processes mediated by the gene product (from Gene Ontology)	45	45745	566	263	.1047
Cellular components where the gene is found (from Gene Ontology)	40	41649	566	298	.1297
UniGene Cluster ID	49	14820	566	419	.1747
Top ten tissue types where gene is expressed	42	42796	566	181	.0824
UniGene Accession No.	44	14820	566	419	.1747
Accession No.s for all representative mRNA sequences in the RefSeq database	46	14820	566	419	.1747
Descriptions associated with the mRNA Accession No.s	2	15556	566	338	.1688

Query expansion using information extracted from online databases failed to improve MAP. When individual queries were examined, some benefited from some kinds of expansion. Addition of identifiers and accession numbers for genes used in the various databases had minimal effect on retrieval, suggesting that these rarely appear in the titles and abstracts of journal articles indexed in MEDLINE. Data categories that consist of larger numbers of text words had mixed results for individual queries but deleterious effects overall. The negative results are probably due to the dilution of terms in the query that match terms in the MEDLINE record. It is possible that some sort of filtering of terms added to ensure greater specificity would improve the results. But before such a filter can be designed successfully, it will probably be necessary to do a detailed failure analysis. Examination of the relevant documents that were not returned by the queries, and of relevant documents returned by the baseline query but lost during query expansion, should provide some insight into what kind of filtering might be successful. It also may be possible to identify some common features that could be exploited during query expansion. In general, terms directly related to gene or protein function appear to have the most promise based on the improvement of individual queries with the addition of data from Gene Ontology or SwissProt.

Table 4 - Number of queries showing improved or worsened mean average precision with each alteration.

Fields Added to Query	Queries Expanded	Queries Improved	Percentage of Queries Improved	Queries Worsened	Percentage of Queries Worsened
Baseline	50	N/A	N/A	N/A	N/A
Chromosome Location	44	4	9.1%	25	56.8%
Locus Link Summary	13	1	7.7%	12	92.3%
SwissProt Accession No.	41	0	0%	0	0%
Protein function (from SwissProt database)	40	9	2.2%	29	72.5%
Relationships to disease (from SwissProt database)	7	1	14.3%	6	85.7%
Molecular functions of the gene product (from Gene Ontology	43	11	25.6%	31	72.1%
Biological processes mediated by the gene product (from Gene Ontology)	45	4	8.9%	39	86.7%
Cellular components where the gene is found (from Gene Ontology)	40	6	15.0%	33	82.5%
UniGene Cluster ID	49	1	2.0%	0	0%
Top ten tissue types where gene is expressed	42	5	11.9%	33	78.6%
UniGene Accession No.	44	0	0%	0	0%
Accession No.s for all representative mRNA sequences in the RefSeq database	46	1	2.2%	1	2.2%
Descriptions associated with the mRNA Accession No.s	2	0	0%	2	100%

The data categories added to queries in these experiments are just a small subset of information that is available about genes. Addition of other types of data, from other databases, might be more successful. For example, greater exploitation of information from Gene Ontology might prove useful. Simply further expanding the queries by including terms from child and parent concepts in the hierarchy is unlikely to improve retrieval, but perhaps the concepts could be useful for filtering the terms from other sources.

One source that was not used in these experiments but might prove useful is Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/Omim/), also available from NCBI. OMIM contains textual summaries about what is known about the role of various genes in human disease. Again, use of the information will probably need to be selective, and possibly undergo filtering. For example, a search on *BRCA1*, a gene related to breast cancer, returns a long summary that includes sections on clinical features of several types of cancers thought to be affected by mutations in this gene, inheritance, clinical management of patients with mutations in this gene, population genetics, gene mapping, molecular genetics, genotype/phenotype correlations, gene function, animal models, allelic variants, and 170 references. OMIM would probably be most useful for a more clinically focused retrieval task than the TREC 2003 Genomics Track task.

One of the features of this task was the generality of the retrieval task, i.e., find all articles about a gene. With minimal constraints on the query, the universe of possible aspects of information is quite large. It is possible that if given a more specific task, query expansion using existing knowledge about a particular aspect of a gene from online databases might make a more positive contribution to the retrieval task. Tailoring the databases used for query expansion to the type of query would be an interesting challenge and perhaps be more likely to produce successful results.

References

1. Evans DA, Lefferts RG, Greffenstette G, Handerson SK, Hersh WR, and Archbold AA. *CLARIT TREC design, experiments, and results*. *The First Text REtrieval Conference (TREC-1)*. 1992. Gaithersburg, MD: National Institute of Standards and Technology. 251-286.
2. Buckley C, Salton G, Allan J, and Singhal A. *Automatic query expansion using SMART: TREC 3*. *Overview of the Third Text REtrieval Conference (TREC-3)*. 1994. Gaithersburg, MD: National Institute of Standards and Technology. 69-80.
3. Srinivasan P, *Query expansion and MEDLINE.* Information Processing and Management, 1996. 32: 431-444.
4. Hersh W, Price S, and Donohoe L. *Assessing thesaurus-based query expansion using the UMLS Metathesaurus*. *Proceedings of the AMIA 2000 Annual Symposium*. 2000. Los Angeles, CA: Hanley & Belfus. 344-348.
5. Wheeler DL, Church DM, Federhen M, Lash AE, Madden TL, Pontius JU, et al., *Database resources of the National Center for Biotechnology.* Nucleic Acids Research, 2003. 31: 28-33.
6. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, et al., *SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.* Nucleic Acids Research, 2003. 31: 219-223.