Edinburgh-Stanford TREC 2003 Genomics Track: Notebook Paper

Miles Osborne[*], Jeffrey Chang[†], Mark Cumiskey[*], Nipun Mehra[†],
Veronica Rotemberg[†], Gail Sinclair[*], Matthew Smillie[*],
Russ B. Altman[†], Bonnie Webber[*]

[*] School of Informatics, University of Edinburgh
{mcumiske,miles,csincla1,bonnie}@inf.ed.ac.uk
m.b.smillie@sms.ed.ac.uk

[†] Department of Genetics, Stanford University
{russ.altman,jeffc,nmehra}@stanford.edu
vmr@smi.stanford.edu

Abstract

We describe our participation in both tasks in the 2003 TREC Genomics track. For the primary task we concentrated mainly upon query expansion and species-specific document searching. An analysis of the variance of possible retrieval results suggested that the official TREC-supplied test set is only a crude approximation of the true system performance. The secondary task we treated as an extraction problem, using a maximum entropy scorer trained on GeneRIF sentences as positives and other sentences as negatives. While our results were not always equivalent to the actual GeneRIFs, on biological grounds many of them appeared better descriptors than the GeneRIFs themselves.

1 Introduction

The School of Informatics at the University of Edinburgh, Stanford's Center for the Study of Language and Information, and the Laboratory for Informatics in the Department of Genetics at Stanford have an on-going collaboration, aimed at improving access to and use of the biomedical literature. In connection with this collaboration, a group of staff and students at the two institutions decided to mount a joint team to participate in the first TREC-Genomics track. Here we outline how we went about dealing with the two tasks.

2 Primary Task: Ad-hoc Document Retrieval

2.1 Partitioning the document set and indexing

Because many species have genes with the same name, we decided to try to eliminate "wrong species" as a source of false positives, by only searching species-specific sub-collections of the original document set (525,938 documents). We produced these species-specific sub-collections (one for each "official" TREC species: *human* (315,714 documents, 60% of the collection), *rat* (31,859 documents, 6% of the collection), *mouse* (28,280 documents, 5.4% of the collection), and *drosophilia* (8,377 documents, 1.6% of the collection), as well as one for *any* "official" TREC species (373,906 documents, 71.1% of the collection)) through a combination of intuition and machine learning. Specifically, we used Rainbow's classification system [1], choosing as classification features, those that were thought to be indicators of species. These included both obvious features (e.g. the MeSH term *Human* as an indicator of human), but also some idiosyncratic ones (e.g. the MeSH term *Predisposition to Disease* also as an indicator of human). Both obvious and idiosyncratic features were selected through careful analysis of the training qrels. Features indicative of other species (e.g.*yeast, frog*) were also used to help classify documents as **not** being about an "official" species. False positives were also avoided by excluding matches against MedLine fields that might contain distracting information (e.g AD - Institutional affiliation and address). In the end, we only considered the title, abstract, MeSH terms and registry number fields of individual documents in creating these species-specific and *all four species* sub-collections.

The 325 relevant documents (qrels) specified in the official TREC training data were used as the training data for species classification, as was an equal number of documents that were about "non-official" species (i.e. *bos taurus*, *danio rerio* and *c. elegans* - obtained through Locuslink). A separate binary classifier was built for each "official" species, as well as one to classify whether a document is about **at least one** of them. To further eliminate possible false positives from the search space, features were added that would allow us to exclude documents that were considered *non-genetic* and therefore unlikely to contain a GeneRIF. The resulting classification accuracy is shown below for a locally-developed held-out test dataset:

Species	**Accuracy (%)**
Human	88.05
Mouse	89.51
Rat	88.80
Drosophila	98.45
All 4	96.46

We also tried using a variety of features and Medline fields to see if we could improve on these results, but none were found that would increase accuracy in the majority of cases.

Another possibility would have been to simply classify species on the basis of their being indexed by the appropriate MeSH term for the species (e.g. query = *Homo Sapiens*: MeSH term = *Human*). While this may have produced a slight increase in classification accuracy (and therefore a slight increase in the document retrieval recall), it would also have substantially increased the search space, which can lead to a substantial decrease in the document retrieval precision. By restricting our search to species-specific sub-collections, we may pay the price in "false negatives" – relevant documents that are missing from the sub-collection because they lack sufficient features to have been included.

The *all 4 species* sub-collection was intended, in part, to deal with this problem in a fall-back strategy, where our initial retrieval using the individual species sets resulted in poor performance. Time, however, denied us the opportunity to research this thoroughly. One technique that was applied was where the first n (500, 700, 800) documents retrieved using the species specific set were retained and the first 1000-n documents retrieved (distinct from the n documents) using the all 4 species document

[1] http://www-2.cs.cmu.edu/˜mccallum/bow/rainbow/

set were tagged on to make up the 1000 documents. This increased our total recall scores. However, this increase was not deemed justifiable for the resultant decrease in precision.

2.2 Query expansion

The recent surge in IR and IE within biological text has highlighted the inconsistency of the terminology within the text. Genes can be referred to in a variety of different ways: the full name, an acronym derived from the full name, by the name of a homologous gene, the resultant product of the gene. Even Locus Link itself is not complete and does not specify all synonymous names of a gene. Conversely, more than one distinct gene can be referred to with the same handle. Thus, for more effective retrieval, we looked at ways to expand our query. Synonyms from genetic databases were sought to complement the set from LocusLink. Analysis of the training queries and their corresponding qrel documents showed other discrepencies within gene symbols. For example, while the same letters may be used, they may differ in case (upper vs lower). Punctuation characters may be added, as well as different representations of the same numeric specifiers (e.g. 2=II=beta). These issues also had to be taken into consideration when formulating our query.

At run-time, the original, TREC-supplied queries were first processed into *unexpanded queries*, with the same format as the queries in Hersh's third preliminary run. Figure 1 shows an example of a TREC-supplied query.

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	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

Figure 1: Original TREC-supplied query

This TREC-supplied query was transformed into the unexpanded query format shown in Figure 2.

> ”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”

Figure 2: An unexpanded query

This unexpanded query uses the Lucene retrieval engine syntax [2]. Terms in quotations must match as a unit, whilst the $^{2.9}$ modifier is an ad-hoc method for boosting the importance of matching against the associated term. The entire set of terms is taken as a disjunction. For this example, since the TREC supplied query indicated that the relevant gene belonged to *homo sapiens*, we used all documents in our *human* sub-collection of the MEDLINE documents when retrieving documents.

We used different query expansion strategies in our two runs: one emphasised precision over recall, and the other tried to recover more documents, with a possible decrease in precision. In the precision-optimised run:

- Parentheticals in the official gene name or a protein product were extracted and appended to the end of the unexpanded query.

- Conjunctions of two words preceding and two words following a hyphen were also appended to query, provided that none of the words were common English. For example, given the query term *CDK-interaction protein 1*, where at least two words follow the hyphen, the term *(”CDK” && ”interaction protein”)* is appended to the query. (&& is Lucene syntax for logical “and”.) Applying this method for punctuation marks other than hyphenation was not found to be successful.

- Acronyms were generated for terms that had more than three words. Again, such acronyms were simply appended to the set of terms.

- Numeric specifiers were removed and the resultant term appended.

- Non-digit specifiers (such as Roman numerals) were converted to digits and vice versa. Again, these extra terms were appended to the end of the query.

- Lowercase and uppercase versions of all terms were also appended.

Our recall-optimised run started with the same query expansion approach as when emphasising precision. In addition though, we retrieved from SwissProt the official gene, associated gene and proteins synonyms for the relevant species. These were appended to the query.

Query expansion of the query shown in Figure 2 results in the much larger query shown in Figure 3, in which the original query terms are followed by upper-case variants, extracted parentheticals, variants with hyphens removed, lower-case variants, etc.

We also tried expanding the query using GO terms, throwing-away common English (non-biological) words, stemming etc, but none of these led to any improvement in recall and/or precision.

After expansion, a query was sent to the Lucene retrieval engine, and all retrieved documents (from a given document sub-collection) were used in our submission. As a cross-check, we also tried the Lemur retrieval engine (http://www.cs.cmu.edu/˜lemur) and obtained similar results. This showed that our system performance was not simply a consequence of using Lucene.

2.3 Results

On the 50-document development set, we obtained the following results using the query expansion strategy that concentrated on obtaining high-precision results (at the expense of possibly not retrieving some documents).

[2] http://jakarta.apache.org/lucene

”cyclin-dependent kinase inhibitor 1A (p21, Cip1)”$^{2.9}$ ”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” ”CYCLIN-DEPENDENT KINASE INHIBITOR 1A (P21, CIP1)”$^{2.9}$ ”CDKN1A”$^{2.9}$ ”P21” ”CIP1” ”SDI1” ”WAF1” ”CAP20” ”CDKN1” ”MDA-6” ”CYCLIN-DEPENDENT KINASE INHIBITOR 1A” ”DNA SYNTHESIS INHIBITOR” ”CDK-INTERACTION PROTEIN 1” ”WILD-TYPE P53-ACTIVATED FRAGMENT 1” ”MELANOMA DIFFERENTIATION ASSOCIATED PROTEIN 6” ”P21, CIP1” ”CYCLINDEPENDENT KINASE INHIBITORA” ”cyclin-dependent kinase inhibitor 1a (p21, cip1)”$^{2.9}$ ”cdkn1a”$^{2.9}$ ”p21” ”cip1” ”sdi1” ”waf1” ”cap20” ”cdkn1” ”mda-6” ”cyclin-dependent kinase inhibitor 1a” ”dna synthesis inhibitor” ”cdk-interaction protein 1” ”wild-type p53-activated fragment 1” ”melanoma differentiation associated protein 6” ”p21, cip1” ”cyclindependent kinase inhibitora” ”CDK 1C” ”CDK 1” ”CIP1” ”WTPAF1” ”MDAP6” ”cyclindependent kinase inhibitorA (p, Cip)” ”cyclindependent kinase inhibitorA” ”cyclindependent kinase inhibitorA” ”cyclindependent kinase inhibitorA” ”CDKinteraction protein” ”wildtype pactivated fragment” ”melanoma differentiation associated protein” ”cyclindependent kinase 1A (p21, Cip1)”$^{2.9}$ ”MDA6” ”cyclindependent kinase 1A” ”CDKinteraction protein 1” ”wildtype p53activated fragment 1” ”cyclin-dependent kinase 1A (p21 Cip1)”$^{2.9}$ ”p21 Cip1” (”1A p21” && ”Cip1”) (”p21 p21” && ”Cip1”)

Figure 3: An expanded query

Retrieved (out of 331)	185 (55.9%)	280 (84.6%)	265 (80.1%)
Average precision	0.3523	0.3939	0.4870
Exact	0.2959	0.3330	0.4711

The first column shows the results without expanding the query; the second column shows the results after expanding the query. The final column shows the results after using the expanded query and the relevant species-specific document sub-collections.

On the official 50-document test set, we obtained the following results, with column labels the same as above:

Retrieved (out of 566)	360 (63.6%)	523 (92.4%)	495 (87.5%)
Average precision	0.1625	0.1870	0.2888
Exact	0.1458	0.1584	0.2636

As can be seen, using our species-specific document sub-collections leads to a significant gain in precision (54.4%), offset by a small drop in recall (5.4%).

We had mostly similar results using our expansion strategy aimed at increasing recall. The development set produced the following results:

Retrieved (out of 331)	185 (55.9%)	331 (100%)	314 (94.9%)
Average precision	0.3523	0.3838	0.4661
Exact	0.2959	0.3342	0.4340

On the official test set, our results were as follows:

Retrieved (out of 566)	360(63.6%)	501(88.5%)	533(94.2%)
Average precision	0.1625	0.1609	0.2690
Exact	0.1458	0.1385	0.2275

The latter figures show that the addition of terms from SwissProt only improved overall performance when applied to the much small species-specific sub-collections. This leads us to conclude that “profligate” query expansion can pay off in the context of “intelligently” targetting the retrieval set.

2.4 (Failed) Attempts at Re-ranking

Information retrieval and question-answering systems often use *re-ranking* techniques, with a view to promoting the rank of relevant documents. We first tried re-ranking the retrieved documents using a maximum entropy model and purely lexico-syntactic features. To give an indication of how well this might work, Hersh’s third set of results (http://medir.ohsu.edu/~genomics/preliminary.html), with a mean average precision of 0.30 could be boosted to a mean precision score of around 0.60. Unfortunately, we could not discover a set of lexico-syntactic features which could capture the notion of relevance when applied to an unseen set of documents and associated queries. We think there may be two reasons for this: First, the relevance of a document to a particular query appears not to be obviously encoded in the document. Secondly, there is considerable variation in how relevant documents correctly match against a query, and no way of obvious general, systematic way to re-rank them.

We then hypothesized that re-ranking documents based on some “function score” would lead to better results. Here we tried three methods, all without success.

First, we attempted to distinguish between articles that talked about function as opposed no function at all. We devised a “Go-iness” score, based on the the number of prominent GO terms that appeared in an article’s abstract. We performed an analysis of GO-terms by first tokenizing each GO term. We then counted GO-tokens in each abstract. Indeed, abstracts with more GO-tokens do tend to discuss biological ”function” more frequently. We checked the distribution of GO terms across abstracts that talk about function and those that do not. We created a training set of function articles by searching Pubmed with the Mesh term ”protein” and for the negative set we searched for clinical and disease abstracts. We then ran a chi-square test of significance over GO-tokens. We weighted more heavily the scores of tokens that were statistically more likely to be found in function articles. While this measure produced a reasonable distinction between function vs non-function articles, it did not create a clear bi-modal distribution, and overall classification performance based on go-iness was not good.

As an alternative, we tried to develop a Maximum Entropy classifier, to distinguish between function and non-function articles. As positive examples, we took articles that were annotated with a molecular function GO term. As negative examples, we downloaded genetics articles without the MeSH term “protein”. While the performance of the resulting classifier was good, precision fell drastically when it was used to re-rank documents, either alone or combined with Lucene scores. As far as we could discern, the lack of improvement was based on the classifier’s inability to distinguish which gene’s function is being talked about. Since Lucene returns documents that talk about more than one gene, highly functional articles discussing the wrong gene were often highly ranked: We found the top-scoring documents do talk about function but not the function of the gene in question.

It also appeared that certain phrases could help distinguish between the function of queried genes and the function of other genes. For example, “on <gene-name>” seemed to be frequently used in documents that were talking about the effects of other genes on the queried gene (but not the function of the queried gene itself). Since we did not have time to devote to parsing, we attempted to use information about word order around a gene name. Since the Maximum Entropy classifier we were using treats each document as a bag of words, we artificially induced ordering by replacing each word that occurred before a gene in a sentence by “XXX<word>” and each that occurred after, by “<word>XXX”. This created two subtypes of <word> for the classifier to use as features. The method still did not improve performance relative to the original ordering produced by Lucene.

Our final attempt at reordering involved a “one gene theory”, based on the observation that an article that talked about many genes seemed less likely to describe a gene’s function, since it was not a focus of the article. For this we ran the gene name identifier by first blacking out the topic gene and checking if an article contains any other genes. We took all sentences that talked only about one gene and tried to create a distinction using maximum entropy. We also experimented with using the title, first sentence that mentions the gene and last sentence to mention the gene. None of our manuevres aided precision. They did, however, inform our approach to the secondary task (Section 3).

2.5 Result Stability

The official test set contained only 50 queries. One could argue that retrieval results using such a small set are unlikely to be a good estimate of the true system performaance.

To test this hypothesis, random sets of 50 TREC-like queries and qrels were formulated from LocusLink and tested on our system for both optimum recall and optimum precision. We tested 150 of these random sets for each expansion method, with results as follows:

	No expansion (%)	**Precision optimised (%)**	**Recall optimised (%)**
Minimum Precision	23.87	29.43	30.65
Maximum Precision	47.63	49.41	45.04
Mean Precision	34.37	37.33	37.43
Standard Deviation	4.29	3.83	3.52
Precision Range	23.76	19.98	14.39
Minimum Recall	41.31	58.48	67.89
Maximum Recall	79.63	90.14	98.28
Mean Recall	61.95	76.94	80.61
Standard Deviation	7.18	5.43	5.94
Recall Range	38.32	31.66	30.39

There are two significant things to note: First is the wide range in precision and recall that can result from a change in 50-element test set. The second is that our precision scores for the actual test set fall significantly below the mean for our random sets, even taking into account the standard deviation. This indicates that, assuming our random sets were representative of typical queries, the TREC test set was a particularly poor query sample (for our system). What we cannot tell from this experiment is whether relative system performance (i.e., comparing systems against one another) is stable or varies with the choice of test set. This would be something worth assessing.

3 Secondary Task: Reproducing GeneRIF Annotation

Using insights from the first task about gene function, we devised two techniques for picking key phrases out of abstracts that are known to be GeneRIFS (Task 2). Both worked well on cursory subjective evaluation by biologists, but did not score well on the DICE measure.

The first method was based on the density of GO terms in a given sentence. Our hypothesis was that a sentence with a very dense concentration of GO terms would be the most likely GeneRIF text. We used the same tokenized GO-terms as described for Task 1. The method can produce good results. Consider for example, a sentence we picked up from Pubmed ID 12080061 for the Locus Link ID 4012. The actual GeneRIF for this document is:

> *Identification of a tankyrase-binding motif in this protein*

The sentence we picked from the same abstract is:

> *TRF1 binding allows tankyrase to regulate telomere dynamics in human cells, whereas IRAP binding presumably allows tankyrase to regulate the targeting of IRAP.*

In many respects, the sentence picked by our method is a better GeneRIF candidate than the actualy GeneRIF itself, thus raising the usual questions about the quality of the gold standard, and the conditions under which GeneRIFs are selected.

The second method we implemented was based on a Maximum Entropy classification of sentences as either GeneRIF or non-GeneRIF. Using a feature selection method based on chi-square distribution of words, we selected features that helped distinguish between sentences with signal (i.e., GeneRIF) and those without. The positive set consisted of all GeneRIFs other than the test set itself. The negative set consisted of all sentences from the GeneRIF containing abstracts other than those sharing more than three non-stop words in common with the GeneRIF. We created a score algorithm for which the output was the sentence with the highest probability of being a GeneRIF.

To determine the probability that a GeneRIF would be found in a particular position, we annotated a set of 200 MedLine entries from LocusLink associated with GeneRIFs. We located the words from the GeneRIF within the title and abstract. In this data set, the GeneRIF came from the title in 74 of 187 cases. So we used 0.4 as probability for the title being taken as the GeneRIF. The GeneRIF was drawn from the final sentence of an abstract in in 49 of 187 cases, and so this probability was set to 0.26. Finally, the GeneRIF came from the first sentence in 3 of the 187 cases. Assuming all other sentences to be equally likely, we uniformly distributed the probabilities on other sentences by dividing by (total sentences, including title - 3)(decomposition of 3 - 1 for title, 1 for first sentence and 1 for last sentence).

In the end, we submitted one run from this method. The final submission was based on a rather ad hoc combination of internal features (using the Maximum Entropy classifier to find GeneRIF-like sentences) and the prior probability that a GeneRIF would be found in the title, first, last or other sentence. We also used the GO-based metric, in which sentences were weighted in terms of their concentration of lexical tokens found in GO terms.

While our methods identified sentences that appeared very reasonable sources of GeneRIFs, our DICE score for Task 2 was not very high. Because our methods stayed at the sentence-level, they suffered when the correct GeneRIF was a specific phrase within the sentence.

4 Discussion

In future work on query expansion, rather than expand all queries equally with the same method, we want to consider a strategic approach to tailor query expansion. This can be achieved by gathering simple statistics for each term looking for occurrences of punctuation used, protein name containing and gene names and vice versa, parentheticals, specifier types etc. So each query can have a unique expansion based on it term constituents.

While GeneRIFs provide a good measure of comparison, we feel that they are neither exhaustive nor the best candidates in many cases. This makes machines unable to objectively match the GeneRIFs. As explained in the example above, some of the senences/phrases that the computer picked up were in fact better at explaining function than the actual GeneRIFs themselves. The other examples are not discussed here but they provide ample evidence to the claim that machines can be used to extract biologically relevant information from text. This can potentially save thousands of man-hours of work. Since the machine works objectively, the criteria for selection of such sentences will objective and standard. A cursory checking by Biologists for most of these candidates would be enough. These methods can also be used to extract GeneRIFs from abstracts before 4/2002 when GeneRIFs were not available.

Acknowledgements

This work has been supported by in part by Scottish Enterprise, through the Stanford-Edinburgh LINK project (Sinclair, Mehra). Jeffrey Chang was funded under NIH grant GM61374, and Veronica Rotemberg, under the Stanford Summer Research program. We would like to thank James Curran for his suggestions and ideas, and Beatrice Alex for annotating the MedLine entries used in developing our approach to Task 2.