MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Enhancing Access to the Bibliome: The TREC Genomics Track

William Hersh, Ravi Teja Bhupatiraju, Sarah Corley

Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Abstract

The growing amount of scientific discovery in genomics and related biomedical disciplines has led to a corresponding increase in the amount of on-line data and information. A new challenge for biomedical researchers has been how to access and manage this ever-increasing quantity of information. The Text Retrieval Conference (TREC) has implemented a Genomics Track to create an experimental environment for research in the use of information retrieval systems in the genomics domain. In the first year of the track, an ad hoc document retrieval task and an information extraction task were carried out by 29 research groups. Future work will focus on more complex data sources, searching tasks, and types of experiments.

Keywords:

Information retrieval, Text Retrieval, genomics, bioinformatics.

Introduction

New genomics and proteomics technologies have had a profound impact on biomedical researchers. Previously, expertise in the sole gene, protein, cellular pathway, etc. of one's research area was enough to carry out productive research. Newer techniques, such as phylogenetic analysis of nucleotide sequences and gene microarrays, have vastly increased the amount of knowledge researchers must master to carry out their work. This problem presents new opportunities for informatics research. We describe a new genomics-based initiative as part of the Text Retrieval Conference (TREC), an initiative to evaluate information retrieval (IR) systems across a variety of domains and tasks.

The "use case" for the TREC Genomics Track is an individual (initially a biomedical researcher but later will be others such as clinicians and/or lay people) who needs to learn about a new scientific area rapidly yet comprehensively. The biomedical researcher may have performed a microarray experiment and discovered that a new gene is involved in the biological process that he or she studies. This gene may have other information associated with it, i.e., it is known to be involved in other disease processes. As a result, the researcher needs to gather and assimilate biomedical literature and other information, using techniques of information retrieval, summarization, and mining.

Text Retrieval Conference

TREC is an annual activity of the IR community aiming to evaluate systems and users. It is sponsored by the National Institute for Standards and Technology (NIST) [1, 2]. The IR field has historically focused on document retrieval, but has expanded in recent years with the growth of new information needs (e.g., question-answering, cross-lingual), data types (e.g., video) and platforms (e.g., the Web). A key feature of TREC is that research groups use a common *test collection* consisting of documents, queries or tasks, and relevance judgments that have determined which documents are relevant to each query or task. The goal is to allow comparisons across systems and approaches in a research-oriented, collegial manner.

TREC is organized into "tracks" of common interest, such as question-answering, multi-lingual retrieval, Web searching, and interactive retrieval. TREC generally works on an annual cycle, with data distributed in the spring, experiments run in the summer, and results presented at the annual conference in the fall. New tracks tend to come about when a critical mass of interest emerges within the community. This was the case for genomics, when IR researchers found themselves increasingly drawn to this domain of relevance to society where rich information resources have already been developed.

Evaluation in TREC is based on the "Cranfield paradigm" that measures system success based on quantities of relevant documents retrieved, in particular the metrics of *recall* and *precision* [3]. Operationally, recall and precision are calculated using a test collection of known documents, queries, and judgments of relevance between them. In most TREC tracks, the two are combined into a single measure of performance, *mean average precision* (MAP). To calculate the MAP for a *run* (i.e., an experiment using a specific IR system and test collection), one first must determine the average precision for each query, which is calculated from the average of all precision values after each relevant document is retrieved. The mean of these average precision values across all queries in the test collection is the MAP.

Some TREC tracks necessitate different evaluation metrics. The Question-Answering track, for example, focuses on finding a single answer to a question as high in the ranked output as possible. As such, the evaluation metric used is *mean reciprocal rank* (MRR) [4]. The performance metric in the Interactive Track has varied depending on the specific user task, but is usually a measure reflective of what the user has been asked to do, such as find one or many answers to a given question [5, 6].

Genomics information resources

Many genomics information resources are available [7]. The best-known of these resources are from the National Center for Biotechnology Information (NCBI), a division of the National

Table 1: GenRIFs for the interleukin 3 (colony-stimulating factor, multiple) gene. This gene was topic 35 in the primary task, is found in the species Homo sapiens, and has LocusLink identifier 3562

PubMed ID	GeneRIF text
11763346	Inhibition of signaling by antisense oligodeoxynucleotides targeting the common beta chain of receptors.
11861295	Ectopically expressed in myeloid leukemic cells with t(5;12)(q31;p13), suggesting that expression of IL3 was deregulated by the translocation, indicating a variant leukemogenic mechanism for translocations involving the 5' end of ETV6.
12002675	Antiapoptotic cytokine IL-3 + SCF + FLT3L influence on proliferation of gamma-irradiated AC133+/CD34+ progenitor cells.
12055233	Monocytes cultured in the presence of IL-3 (plus IL-4) differentiate into dendritic cells that produce less IL-12 and shift T helper (Th) cell responses toward a Th2 cytokine pattern.
12093816	Data suggest that increased activity of mutated interleukin 3 is due to a change from a rare ligand to a common one, allowing the increase in IL-3-dependent signaling.
12135758	Role in potentiating hematopoietic cell migration.
12165512	The IL-3 gene is regulated by two enhancers that have distinct but overlapping tissue specificities.

Library of Medicine (NLM) that maintains most of the NLM's genomics-related databases [8]. Key features of NCBI data include linkage and annotation. Linkage among resources allows the user to explore different types of knowledge across resources. For example, the original research documenting the discovery of a gene function appears in MEDLINE (the bibliographic database of medical literature, accessed by PubMed and other systems), with links to the nucleotide sequence in GenBank, the structure of the protein in the Molecular Modeling Database (MMDB), and an overview of the diseases it may cause in humans in the Online Mendelian Inheritance in Man (OMIM) textbook. LocusLink serves as a switchboard to integrate these resources together as well as provide annotation of the gene's function using the widely accepted GeneOntology (GO) [9].

Development of the Genomics Track

The TREC Genomics Track began with exploratory, consensus-building deliberations. The first activity was a Web survey soliciting ideas for the track in early 2002. Over 80 individuals responded, revealing diverse interests in IR and information extraction (IE) tasks, but clustered around three areas: extraction of knowledge from databases, automating the annotation of genes and proteins, and retrieval across heterogeneous databases. Respondents from the IR community expressed the most enthusiasm for the latter task. All respondents were interested in using public databases, mainly those from the NCBI.

Activity also consisted of three workshops, held at the Joint Conference on Digital Libraries (JCDL) 2002, TREC 2002, and the Pacific Symposium on Biocomputing (PSB) 2003. These workshops led to the plan for the first year of the track, which took place in 2003. For the 2003 track, decisions for data, queries, and relevance judgments were made in the context of having few resources for data acquisition and relevance judgments. As such, the choice of which queries and documents to use was guided by the availability of existing resources that could avoid the need for labor-intensive (i.e., costly) relevance judgments. A consensus emerged that two experimental tasks would be developed: an ad hoc retrieval task and an extraction task. Both tasks were based on the availability of Gene Reference into Function (GeneRIF) data in the NCBI LocusLink database [10]. Each GeneRIF entry consists of a statement about the function of a gene along with a pointer to the MEDLINE reference for the article that discovered that data [11]. This would allow us to have proxies for relevance judgments in the retrieval task and targets for designating important text in the extraction task. Table 1 shows an example of GeneRIFs for a specific gene which was used in the primary task described below.

The ad hoc task was the primary task. A preliminary analysis in January, 2003 identified nearly 7,000 genes with one or more GeneRIFs. There were 246 genes with 10 or more GeneRIFs. We randomly selected genes to use as topics from the entire spectrum of numbers of GeneRIFs. Based on past IR work, which has shown that the "stability" of recall-precision numbers in batch retrieval experiments requires at least 25 and ideally 50 topics [12], we decided to use 50 genes as topics. We selected another 50 genes as "training" topics where we also provided the GeneRIF (relevance) data to researchers to develop their systems for the official "test" topics where the relevance data was not provided until the results were submitted to NIST.

The secondary task for 2003 was an exploratory task of extracting the GeneRIF statement given the gene name and pointer to the MEDLINE record. Research groups were provided both, with lexical overlap of the GeneRIF statement as measured by the Dice coefficient and some variations of it serving as the measures for success. The full text of the articles was also acquired through Highwire Press (www.highwire.org), which publishes the full text of over 300 biomedical journals. (Highwire has served as a intermediary to help research groups such as ours obtain journal data for their work.)

Primary Task

The primary task for 2003, an ad hoc document retrieval, required a document collection, topics, and relevance judgments.

Documents

The document collection consisted of 525,938 MEDLINE records where indexing was completed between 4/1/2002 and 4/1/2003. The MEDLINE records were provided in a single compressed text file using the standard NLM MEDLINE format (although an XML version was also available).

Table 2: Names for the interleukin 3 (colony-stimulating factor, multiple) gene from LocusLink

OFFICIAL_GENE_NAME	interleukin 3 (colony-stimulating factor, multiple)
OFFICIAL_SYMBOL	IL3
ALIAS_SYMBOL	IL-3
ALIAS_SYMBOL	MCGF
ALIAS_SYMBOL	MULTI-CSF
PREFERRED_PRODUCT	interleukin 3 precursor
PRODUCT	interleukin 3 precursor
ALIAS_PROT	mast-cell growth factor
ALIAS_PROT	P-cell stimulating factor
ALIAS_PROT	hematopoietic growth factor
ALIAS_PROT	multilineage-colony-stimulating factor

Table 3: Top five, median, gene names, and lowest performing runs in primary task for MAP, mean relevant at 10 documents retrieved, and mean relevant at 20 documents retrieved.

Organization or designation	**MAP**	**Relevant @ 10**	**Relevant @ 20**
National Library of Medicine #1	0.4165	3.16	4.84
National Library of Medicine #2	0.3994	3.20	4.56
National Research Council #1	0.3941	2.94	4.38
University of California Berkeley	0.3912	3.06	4.46
National Research Council #2	0.3771	2.76	4.36
Median	0.2001	1.50	2.44
Gene names	0.1372	1.18	0.88
Lowest	0.0271	0.22	0.60

Topics

The topics consisted of gene names, with the specific task being as follows: *For gene X, find all MEDLINE references that focus on the basic biology of the gene or its protein products from the designated organism. Basic biology includes isolation, structure, genetics and function of genes/proteins in normal and disease states.* Each gene was also associated with one of four possible organisms: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, and *Drosophila melanogaster*. Table 2 shows an example of a gene name and its synonyms from LocusLink.

We distributed training and test topic sets of 50 genes each. The training data were provided for groups to become familiar with the data and tune their systems for it. For each set of topics, we randomly chose gene names that were distributed across the spectrum of organisms, the number of GeneRIFs (many to few), or whether or not the gene names are Medical Subject Heading (MeSH) indexing terms. Training data included a file of GeneRIFs that comprised the relevance judgments for those topics.

Methods

Track participants were not allowed to use GeneRIF data to augment their queries. While we recognized that GeneRIFs were, like the rest of LocusLink, publicly available, we worked on the *honor system* of not using GeneRIF data. Recall and precision were calculated using MAP in the standard TREC approach of participants submitting their results in the format for input to the trec_eval program (code available at ftp.cs.cornell.edu/pub/smart/). The trec_eval program requires two files for input. One file is the topic-document output, sorted by each topic and then subsorted by the order of the IR system output for a given topic. The second file required for trec_eval is the relevance judgments, which are called *qrels* in TREC jargon. (More information about qrels can be found at trec.nist.gov/data/qrels_eng/.)

Results

A total of 49 official runs were submitted by 25 different research groups. Table 3 shows the results from the top five runs as well as the run ranking at the median of all results, a run using only the official gene name as the query, and the lowest ranking run. (Statistical analysis in the form of an ANOVA with posthoc pairwise comparisons will be available by November, 2003 but was not ready by the deadline for this paper.) Table 4 demonstrates the wide variation in results for different research groups, showing the best, median, and worst values for the number of relevant documents retrieved in the top 10 and top 20 documents and the MAP for topic 35, which is the gene used in Tables 1, 2.

The top-ranking runs came from a research group at the National Library of Medicine [13].

Table 4: Distribution of results for topic 35, which had 7 GeneRIFs, for MAP, mean relevant at 10 documents retrieved, and mean relevant at 20 documents retrieved

MAP			Relevant @ 10			Relevant @ 20		
Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
0.4136	0.0647	0	6	1	0	4	1	0

Table 5: Relevance analysis for ten topics from the training data

Topic Number	GeneRIFs-Total	GeneRIFs-Relevant	Retrieved-Total	Retrieved-Not Relevant	Retrieved-Relevant and a GRIF	Retrieved-Relevant and not a GRIF	Retrieved-Relevant and another species
2	2	2	20	18	1	0	1
3	11	11	20	18	1	1	0
4	8	8	20	0	3	11	6
10	11	11	20	2	0	18	0
11	17	17	20	0	6	14	0
28	7	7	20	0	1	19	0
35	1	1	20	20	0	0	0
36	5	5	20	1	1	9	10
40	8	8	20	5	4	9	2
48	4	4	20	6	4	4	6
Total	74	74	200	70 (35%)	21 (10.5%)	85 (42.5%)	25 (12.5%)

This group used a variety of domain-specific techniques for identifying gene names along with ranking techniques employing weighting for the relative importance of features identified from documents. Their second run employed additional rules for using Medical Subject Headings (MeSH) terms and a Bayesian classifier for additional weighting. The next-highest ranking runs came from the National Research Council of Canada, which oriented its system toward achieving very high recall while minimizing the total number of documents retrieved, and the University of California Berkeley, which also employed domain-specific techniques for recognizing gene names and a Bayesian model [14].

We also performed an analysis of relevance to determine how well GeneRIFs covered the relevant documents. One of us (SC) assessed the top 20 documents retrieved as well as all GeneRIFs for 10 queries, determining if the GeneRIFs were indeed relevant in terms of the retrieval task as well as assessing how many articles were relevant but not designated as GeneRIFs. As shown in Table 5, this analysis found that an article pointed to by a GeneRIF was always relevant in the classic IR sense but that there were many "false negatives," i.e., articles that were relevant but do not have a GeneRIF designation.

Secondary task

The goal of the secondary task was to reproduce the GeneRIF annotation. One possibility for measuring success would be to calculate some sort of overlap measure between words that research groups nominate for annotation with those actually selected by NLM. A problem, however, is that while some GeneRIF snippets are direct quotations from article abstracts, others are paraphrased. Furthermore, there can be other legitimate references to basic gene biology beyond the official GeneRIF snippet. An analysis of several thousand GeneRIFs found that 95% of GeneRIF snippets contained some text from the title or abstract of the article (personal communication, James Mork and Alan Aronson). About 42% of the matches were direct "cut and paste" from the title or abstract, 25% contained significant runs of words from pieces of the title or abstract. The goal of the secondary task was to reproduce the GeneRIF from the MEDLINE record. Because of the exploratory nature of this task, we did not provide any training data. Groups were asked to use automated approaches and describe them frankly in their reports.

Data

The data for the secondary task consisted of 139 GeneRIFs representing all of the articles appearing in five journals (*Journal of Biological Chemistry*, *Journal of Cell Biology*, *Nucleic Acids Research*, *Proceedings of the National Acad. of Sciences*, and *Science*) during the latter half of 2002. We obtained the full text of these articles from Highwire Press (www.highwire.org), who obtained permission for our use of them from the publishers.

Methods

The original plan for assessing the secondary task was to use the Dice coefficient, which measures overlap of two strings. That is, the overlap between the candidate GeneRIF and actual GeneRIF was calculated. For two strings A and B, define X as the number of words in A, Y as the number of words in B, and Z as the number of words occurring in both A and B.

The Dice coefficient is therefore measured as follows:

$$\text{Dice}(A, B) = (2 * Z)/(X + Y)$$

It quickly became apparent that this measure was quite limited. It did not, for example, perform any "normalization" of words, such as stop word removal or stemming. It also did not give any

credit for words occurring more than once in both strings. Finally, it assumed the strings were simply bags of words and did account for word order or phrases. A consensus of track participants developed four derivatives of the classic Dice measurement for the task:

- Classic Dice with stop words and stemming - The basic measure is the classic Dice formula using a common stop word list and the Porter stemming algorithm.
- Modified Unigram Dice - The next measure gives added weight to terms that occur multiple times in both strings. In particular, each set of words in a string is multi-set, with the number of co-occurring words measured by the minimum number of co-occurrences.
- Bigram Dice - This measure gives some additional weight to proper word order. Instead of measuring the Dice coefficient on single words, it measures it on bigrams.
- Bigram Phrases - Bigrams do not always represent legitimate phrases. Stop words like articles and prepositions sometimes occur between content words such that straight bigrams of content words do not represent legitimate phrases. A further measure therefore only includes bigrams that have not had intervening stop words filtered.

Results

A total of 25 runs were submitted by 14 research groups. Ten of these groups participated in the primary task while four took part only in this task. (As of mid-September, results for this task have been submitted but not officially scored. As with the primary task, complete analysis of results will be completed by mid-November, 2003.)

Future Directions

In its first year of existence, the TREC Genomics Track drew a great deal of interest from both computer science groups interested in the effectiveness of IR algorithms in a new domain as well as bioinformatics groups interested in methods for solving information problems in the domain of their work. A five-year plan for expanding the track into new content types (e.g., full-text journal articles, textbooks, and other resources), topic types (e.g., answering specific questions), interactive experiments, and different users (e.g., clinicians, consumers) has been funded by a recent grant from the US National Science Foundation. The interest expressed in the first year of the track along with the concrete plan for further development indicates this effort should contribute to important research findings applying IR techniques to genomics problems.

References

[1] Voorhees EM and Harman D, Overview of the Sixth Text REtrieval Conference (TREC). *Information Processing and Management*, 2000. 36: 3-36.

[2] Voorhees E and Harman D. Overview of TREC 2001. *Proceedings of the Text Retrieval Conference 2001*. 2001. Gaithersburg, MD. 1-15.

[3] Hersh WR, *Information Retrieval: A Health and Biomedical Perspective* (Second Edition). 2003, New York: Springer-Verlag.

[4] Voorhees EM and Tice DM. Building a question answering test collection. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2000. Athens, Greece: ACM Press. 200-207.

[5] Hersh WR, Interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management*, 2001. 37: 365-366.

[6] Hersh W, et al., Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing and Management,* 2001. 37: 383-402.

[7] Baxevanis A, The Molecular Biology Database Collection: 2003 update. Nucleic Acids Research, 2003. 31: 1-12.

[8] Wheeler DL, et al., *Database resources of the National Center for Biotechnology. Nucleic Acids Research*, 2003. 31: 28-33.

[9] Ashburner M, et al., Gene Ontology: tool for the unification of biology. Nature Genetics, 2000. 25: 25-29.

[10] Pruitt KD and Maglott DR, RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 2002. 29: 137-140.

[11] Mitchell JA, et al. Gene indexing: characterization and analysis of NLM's GeneRIFs. *Proceedings of the AMIA 2003 Annual Symposium.* 2003. Washington, DC: Hanley & Belfus. in press.

[12] Buckley C and Voorhees E. Evaluating evaluation measure stability. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2000. Athens, Greece: ACM Press. 33-40.

[13] Aronson AR. Methods for accurate retrieval of MEDLINE citations in genomics. The Twelfth Text Retrieval Conference (TREC 2003). 2003. Gaithersburg, MD: National Institute of Standards and Technology. in press.

[14] Hearst M. UC Berkeley BioText Group. The Twelfth Text Retrieval Conference (TREC 2003). 2003. Gaithersburg, MD: National Institute of Standards and Technology. in press.

Address for correspondence

William Hersh, M.D.
Dept of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
3181 SW Sam Jackson Park Rd., Mail Code BICC
Portland, OR, USA 97239
Email: hersh@ohsu.edu