TREC Genomics Track Overview

William Hersh, Ravi Teja Bhupatiraju
Oregon Health & Science University
Portland, OR, USA
{hersh,bhupatir}@ohsu.edu

The first year of TREC Genomics Track featured two tasks: ad hoc retrieval and information extraction. Both tasks centered around the Gene Reference into Function (GeneRIF) resource of the National Library of Medicine, which was used as both pseudorelevance judgments for ad hoc document retrieval as well as target text for information extraction. The track attracted 29 groups who participated in one or both tasks.

The growing amount of scientific discovery in genomics and related biomedical disciplines has led to a corresponding growth in the amount of on-line data and information. A growing challenge for biomedical researchers is how to access and manage this ever-increasing quantity of information. This situation presents opportunities and challenges for the information retrieval (IR) field. IR has historically focused on document retrieval, but the field 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). This paper describes the events leading up to the first year of TREC Genomics Track, the first year's results, and future directions for subsequent years.

Genomics and Information Resources

The field of *genomics* is concerned with the *genome*, which is usually defined as the genetic material of living organisms. Its research focuses on the *central dogma* of biology: deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA), which serves to translate the nucleotide sequences of DNA into proteins. The latter are responsible for functions in living organisms and the collection of all proteins in is increasingly called the *proteome*. With the advent of new technologies for sequencing the genome and proteome, along with other tools for identifying the expression of genes, structures of proteins, and so forth, the face of biological research has become increasingly data-intensive, creating great challenges for scientists who formerly dealt with relatively modest amounts of data in their research.

The growth of biological data has resulted in a correspondingly large increase in scientific knowledge in what biologists sometimes call the *bibliome* or literature of biology. A great deal of biological information resources have become available in recent years (Baxevanis, 2003). Probably the most important of these are from 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) that maintains most of the NLM's genomics-related databases (Wheeler, Church et al., 2003).

Key features of NCBI resources 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 (http://www.ncbi.nlm.nih.gov/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, www.geneontology.org). Genes with known locations are also into maps which denote their locations of genes on chromosomes. PubMed also provides linkages to full-text journal articles on the Web sites of publishers.

Additional genomics resources exist beyond the NCBI. Of particular note are the model organism genome databases, such as:

Mouse Genome Informatics - www.informatics.jax.org
Saccharomyces Genome Database - http://genome-www.stanford.edu/Saccharomyces/
Flybase Database of the Drosophilia Genome - flybase.bio.indiana.edu

As with the NCBI resources, these resources provide rich linkage and annotation.

Preliminary Activities of the Genomics Track

The genesis of the Genomics Track came in 2001, when interest was expressed by the TREC Program Committee for moving into new types of data, including types which were more structured than the usual document collections from newswire. A number of discussions among interested people led to the first activity of the track, which was a Web survey soliciting ideas that took place 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, automated or semi-automated 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 was hoped would take place in 2003. A significant constraint on the track was lack of resources; i.e., NIST did not have the in-house resources to obtain documents or perform relevance judgments in this domain. As such, the choice of tasks, queries, documents, etc. would need to be guided by the availability of existing resources, including something that could be used to serve as proxies for relevance judgments. Fortunately, the track identified a valuable resource from NCBI: Gene Reference into Function (GeneRIF) data in the LocusLink database. 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 (see Table 1).

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. As past IR work has shown that the "stability" of recall-precision numbers in batch retrieval experiments requires at least 25 and ideally 50 topics (Buckley and Voorhees, 2000), this would provide ample data for experiments.

The workshops also resulted in a use case guiding the first year's experiments, which was the biological researcher or graduate student (i.e., someone who already has considerable general domain knowledge) who is confronted with the need to learn about a new scientific area quickly. Perhaps he or she has performed a gene expression array experiment identifying genes not previously known to be involved in the biological process he or she has been investigating. Now he or she must get up to speed quickly with knowledge of these genes.

The GeneRIFs allowed the track to pursue two tasks satisfying the interests of a larger audience: an ad hoc retrieval task and an IE task. The ad hoc task was designated the primary task and was structured very similar to most previous TREC ad hoc tasks (e.g., ad hoc tasks of TREC 1-10, Web track, etc.). It was recognized that GeneRIFs could serve as pseudorelevance judgments, even though it was suspected (and later verified, see below) that they were incomplete from that standpoint.

GeneRIFs could also be used as targets for IE, and this was chosen to be the secondary task. The secondary task was more exploratory in nature: extracting the GeneRIF statement from the MEDLINE record or the article proper.

Table 1 - GeneRIFs for the gene *Interleukin 3 (colony-stimulating factor, multiple)* from LocusLink. The PubMed ID and citation are from the MEDLINE database.

LocusLink ID	PubMed ID	Citation	GeneRIF text
3562	11763346	Antisense Nucleic Acid Drug Dev 2001 Oct;11(5):289-300.	inhibition of signaling by antisense oligodeoxynucleotides targeting the common beta chain of receptors
3562	11861295	Blood 2002 Mar 1;99(5):1776-84.	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
3562	12002675	Folia Biol (Praha) 2002;48(2):51-7.	Antiapoptotic cytokine IL-3 + SCF + FLT3L influence on proliferation of gamma-irradiated AC133+/CD34+ progenitor cells.
3562	12055233	J Immunol 2002 Jun 15;168(12):6199-207.	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.
3562	12093816	J Biol Chem 2002 Oct 11;277(41):38764-71.	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.
3562	12135758	FEBS Lett 2002 Jul 31;524(1-3):149-53.	role in potentiating hematopoietic cell migration
3562	12165512	J Immunol 2002 Aug 15;169(4):1876-86.	The IL-3 gene is regulated by two enhancers that have distinct but overlapping tissue specificities.

Research groups were charged with maximizing the lexical overlap of the GeneRIF statement as measured by the Dice coefficient and some derivatives of it. Full-text articles were provided through Highwire Press (www.highwire.org), which publishes the full text of over 400 biomedical journals. Highwire does not own the copyrights to the journals, but has served as an intermediary to help various IR and other research groups obtain journal data for their work. Highwire facilitated interaction with publishers to obtain content for experiments.

Primary task

As noted above, the primary task for 2003 consisted of ad hoc document retrieval. This type of task requires 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 the standard NLM MEDLINE format (although an XML version was available). The fields were indicated by their 2-3 letter abbreviation. The fields likely to be most important to track participants were: PubMed Unique Identifier (PMID), title (TI), abstract (AB), and MeSH headings (MH). A description of all the fields in a MEDLINE record can be found in the PubMed help file at: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#MEDLINEDisplayFormat

Topics

The topics consisted of gene names, with the specific task being deriving from the definition of a GeneRIF (Mitchell, Aronson et al., 2003):

> 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.

We distributed training and test topic sets of 50 genes each. The training data were distributed first, allowing groups to get an idea of what the data in the track were like and tune their systems. The test data were the topics for the official runs in the track. For each set of 50 topics, we randomly chose gene names that were distributed across the spectrum of organisms, the number of GeneRIFs (many to few), name types (see below), and whether or not the gene names were Medical Subject Heading (MeSH) indexing terms. We also distributed the GeneRIFs for all of the training topics to allow groups to see the targets of their systems' retrieval efforts. GeneRIFs for the test topics were not distributed until after the deadline for the submission of official results.

Gene Names

LocusLink also contains a variety of names for each gene. Many researchers have lamented the pervasiveness of synonymy and polysemy in gene naming (O'Neill, 2003). Table 2 shows the multiple gene names for the *Interleukin 3 (colony-stimulating factor, multiple)* gene whose GeneRIFs were shown in Table 1.

Although many genes are present across multiple organisms (e.g., humans, mice, and rats produce and utilize insulin), LocusLink maintains a separate record for each gene in a given species. We chose to limit genes to four possible organisms:

Homo sapiens - human
Mus musculus - mouse
Rattus norvegicus - rat
Drosophila melanogaster - fruit fly

Relevance Judgments

For reasons described above, the relevance judgments for the 2003 track consisted of GeneRIFs. 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 research groups not using GeneRIF data.

Table 2 - Names for the gene *Interleukin 3 (colony-stimulating factor, multiple)* from LocusLink.

LocusLink ID	Organism	Gene name type	Gene name
3562	Homo sapiens	OFFICIAL_GENE_NAME	interleukin 3 (colony-stimulating factor, multiple)
3562	Homo sapiens	OFFICIAL_SYMBOL	IL3
3562	Homo sapiens	ALIAS_SYMBOL	IL-3
3562	Homo sapiens	ALIAS_SYMBOL	MCGF
3562	Homo sapiens	ALIAS_SYMBOL	MULTI-CSF
3562	Homo sapiens	PREFERRED_PRODUCT	interleukin 3 precursor
3562	Homo sapiens	PRODUCT	interleukin 3 precursor
3562	Homo sapiens	ALIAS_PROT	mast-cell growth factor
3562	Homo sapiens	ALIAS_PROT	P-cell stimulating factor
3562	Homo sapiens	ALIAS_PROT	hematopoietic growth factor
3562	Homo sapiens	ALIAS_PROT	multilineage-colony-stimulating factor

We calculated recall and precision in the classic IR way, using the preferred TREC statistic of *mean average precision* (average precision at each point a relevant document is retrieved, also called MAP). This was done in the standard TREC fashion of participants submitting their results in the format for input to the trec_eval program. (Groups were directed to the repository of code for trec_eval at ftp://ftp.cs.cornell.edu/pub/smart/. There are several versions of trec_eval, which differ mainly in the additional statistics they calculate in their output.) 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 http://trec.nist.gov/data/qrels_eng/).

Training Data Runs

The training data runs not only allowed groups to become familiar with the data, but also allowed for discovery of some "quirks" with the training data qrels:

- A number of qrels represented documents not present in the document collection.
- Three topics had no qrels in the document collection: 21, 35, and 49.

This enabled us to make sure these problems did not exist before finalizing the test data. Due to the unstable nature of recall-precision for topics with very small numbers of qrels, we made the decision to use only gene names that had a minimum of three qrels in the collection for the test topics.

We also performed an analysis of relevance for 10 queries from the training data. This was done by manually judging relevance for all GeneRIFs as well as all other documents in the top 20 retrieved by the best OHSU training data run (or all documents if less than 20 retrieved). The relevance judgments were performed by an individual with a medical background enrolled in the OHSU medical informatics graduate program who had taken a course in IR. This analysis validated our *a priori* assumptions that all articles pointed to by GeneRIFs were relevant in the classic IR sense and that there were many "false negatives" (i.e., articles that were relevant but did not have a GeneRIF designation). We also discovered another phenomenon: documents that were relevant but for the gene in a species other than that designated by the LocusLink record. In the analysis of the top 20 ranking documents retrieved from the best OHSU training run, we found that:

- 35.0% of documents retrieved were not relevant
- 10.5% of documents retrieved were relevant and were GeneRIFs
- 42.5% of documents retrieved were relevant and not GeneRIFs
- 12.5% of documents retrieved were relevant and from a different species

Official Runs

A total of 25 groups submitted 49 official runs for scoring. Table 3 lists the results for each run, consisting of the run tag, whether the run was purely automatic or used some manual processing, and three results: MAP, number relevant at 10 documents retrieved, and number relevant at 20 documents retrieved. The final two rows of the table show the mean and median for each result. An analysis of variance with posthoc pairwise comparisons will be reported in a subsequent paper.

The run with the highest MAP was NLMUMDSE, with a mean MAP of 0.4165. This and the next-highest performing run came from an NLM-based research group (not affiliated with the operations of the library) (Kayaalp, Aronson et al., 2003). They used a search engine developed for the ClinicalTrials.gov database. They achieved good results from:

- Identifying species through use of MeSH terms and other simple rules
- Recognizing terms or their synonyms or lexical variants in non-text fields, in particular MeSH and substance name (RN)
- Using additional general key words, such as genetics, sequence, etc.

A second run with a system that added MeSH terms and other controlled vocabulary along with collocation networks did not improve performance with this data.

Runs from UC Berkeley (Bhalotia, Nakov et al., 2003) and the National Research Council of Canada (deBruin and Martin, 2003) ranked next highest. Both of their approaches benefited from rules for recognizing gene name synonyms and filtering for organism name. The UC Berkeley approach included a machine learning algorithm to classify documents likely to have GeneRIFs assigned to them and document ranking based on gene name occurrence rules. The NRC approach added unsupervised relevance feedback to find additional relevant articles and ranking based on TF*IDF query term weighting.

The Waterloo group also did well, using what could be best described as "database-specific" (as opposed to "domain-specific") techniques that included (Yeung, Clarke et al., 2003):

- Query formulation using fusion of Okapi weighting plus handling of punctuation plus pluralization as well as gene name bigrams
- Recognition of gene name in substance name field
- Query expansion on relevant substance names

It was apparent from the above groups that searching in the MeSH and substance name fields, along with filtering for species, accounted for the best performance. At least two other groups also found substantial benefit from organism name filtering, the National Research Council of Canada (deBruin and Martin, 2003) and Tarragon Consulting (Tong, Quackenbush et al., 2003). No groups attempted to model gene "function" in the sense of the GeneRIFs.

Approaches that used standard IR techniques shown to work best with traditional TREC data (i.e., newswire) performed less well. The Neuchatel group tried many permutations of advanced features from SMART (Savoy, Rasolofo et al., 2003). They obtained their best results with Okapi weighting, pivoted normalization, and query expansion, but they fell near the median of all groups. Likewise, the Illinois-UC group used a variant of language modeling and also performed near the median (Zhai, Tao et al., 2003).

As with many TREC experiments over the years, variation across topics and even within them across groups was substantial. Table 4 shows the variation of results for topic 35, the gene whose GeneRIFs and gene names were displayed in earlier tables.

We also carried out a relevance similar to that described for the training data with all 50 test topics. Again, all GeneRIFs as well as the top 20 documents retrieved (or all documents if less than 20 retrieved) in the best OHSU run (ohsuboost) were analyzed by the individual described above. Once again, we found that virtually all GeneRIFs were relevant (551/566, 97.3%), although a small number were relevant in other species (13/566, 2.3%) or indeterminate because the abstract was not accessible in MEDLINE to allow judgment (2/566, 0.4%). However, we also found again that substantial numbers of documents deemed relevant by our judge were not designated as GeneRIFs. Table 5 summarizes the analysis of retrieved documents.

Secondary Task

There is much interest in the bioinformatics community in IE. This comes in part from the desire to allow scientists to learn about new topics as quickly as they can, preferably without having to read and synthesize many papers. The specific task was to reproduce the GeneRIF annotation. As this task was more exploratory in nature and had an uncertain "gold standard," groups were instructed to attempt the task and compare their methods and results. Because of the exploratory nature of the secondary task, we did not provide any training data.

Table 3 - Official primary task runs, sorted by mean average precision.

Run Tag	Run Type	Mean Average Precision	Relevant @ 10 documents retrieved	Relevant @ 20 documents retrieved
NLMUMDSE	automatic	0.4165	3.16	4.84
NLMUMDSRB	manual	0.3994	3.20	4.56
nrc1	automatic	0.3941	2.94	4.38
biotext1	automatic	0.3912	3.06	4.46
nrc2	automatic	0.3771	2.76	4.36
biotext0	automatic	0.3753	2.92	4.30
uwmtg03btrf	automatic	0.3534	2.28	3.68
uwmtg03atrf	automatic	0.3479	2.48	4.00
axon2	automatic	0.3173	2.50	3.86
axon1	automatic	0.3118	2.40	3.78
CSUSM2	automatic	0.3079	2.68	3.76
edstanrecall	automatic	0.3015	2.60	3.74
edstanprec	automatic	0.2984	2.60	3.74
KUBIOIRNE	automatic	0.2980	2.32	3.42
KUBIOIRRAW	automatic	0.2937	2.24	3.38
CSUSM1	automatic	0.2859	2.56	3.52
tgnBaseline	manual	0.2837	2.18	3.52
IBMbt1	automatic	0.2823	2.26	3.32
tgnVariant1	manual	0.2791	2.22	3.56
aoyama	automatic	0.2277	1.90	2.92
aoyama2	automatic	0.2276	1.92	2.92
IBMbt2	automatic	0.2259	1.80	2.84
UIowaGN1	automatic	0.2064	2.02	3.40
UIUC03Gb	automatic	0.2001	1.50	2.44
SCAI	automatic	0.1960	1.42	2.60
utafil	manual	0.1931	1.48	2.40
utaband	manual	0.1927	1.54	2.62
UIUC03Ga	automatic	0.1925	1.58	2.32
UBgenomeBGNE	automatic	0.1867	1.44	2.14
UniNEg1	automatic	0.1852	1.28	2.12
humG03ns	automatic	0.1847	1.58	2.46
UniNEg2	automatic	0.1802	1.30	2.10
ErasmusMC3	automatic	0.1770	1.36	2.28
ErasmusMC2	automatic	0.1754	1.38	2.32
humG03ns5	automatic	0.1753	1.48	2.34
ohsuboost	automatic	0.1747	1.58	2.36
DcuMesh1	automatic	0.1669	1.36	2.08
DcuMesh2	automatic	0.1667	1.36	1.96
dayrutgers1	automatic	0.1652	1.34	2.40
dayrutgers2	automatic	0.1636	1.32	2.06
UniNEg5	automatic	0.1635	1.28	2.00
UniNEg4	automatic	0.1623	1.30	2.10
balsc3	automatic	0.1528	1.44	2.10
UBgenomRFB1	automatic	0.1511	1.16	1.84
UBgenomRFB2	automatic	0.1493	1.12	1.80
balsc2	automatic	0.1481	1.36	2.34
StreamSage3	automatic	0.0508	0.70	0.80
StreamSage4	automatic	0.0508	0.70	0.80
vvP05mil3	automatic	0.0271	0.22	0.60
Mean		0.2313	1.85	2.85
Median		0.1960	1.58	2.60

Table 4 - Best, median, and worst scores for the topic, *Interleukin 3 (colony-stimulating factor, multiple)*.

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

Table 5 - Classification of relevance of retrieved documents from best OHSU run organized by whether document, for a given query, is or is not a GeneRIF and is relevant, not relevant, relevant in another species, or unable to be judged due to no abstract in MEDLINE record.

GeneRIF	And	Number	Percentage
GeneRIF	Relevant	117	12.7%
GeneRIF	Not relevant	0	0.0%
GeneRIF	Relevant in another species	2	0.2%
GeneRIF	No abstract (unable to judge)	0	0.0%
Not a GeneRIF	Relevant	386	41.8%
Not a GeneRIF	Not relevant	85	9.2%
Not a GeneRIF	Relevant in another species	333	36.1%
Not a GeneRIF	No abstract (unable to judge)	0	0.0%
Total		923	100.0%

Consensus discussions yielded the notion that measuring success would be best calculated by some sort of overlap measure between words nominated for annotation and those actually selected in the GeneRIF. A problem, however, was that while some GeneRIF snippets were direct quotations from article abstracts, others were paraphrased. Furthermore, there were other legitimate references to basic gene biology beyond the official GeneRIF snippet. A preliminary analysis by Jim Mork and Lan Aronson of NLM found that 95% of GeneRIF snippets contained some text from the title or abstract of the article. About 42% of the matches were direct "cut and paste" from the title or abstract, and another 25% contained significant runs of words from pieces of the title or abstract.

Data

The data for the secondary task consisted of 139 GeneRIFs representing all of the articles appearing in five journals for which we could obtain full text from Highwire (*Journal of Biological Chemistry*, *Journal of Cell Biology*, *Nucleic Acids Research*, *Proceedings of the National Academy of Sciences*, and *Science*) that were published during the latter half of 2002.

Performance Measures

The original plan for assessing the secondary task was to use the Dice coefficient, which measures overlap of two strings. In this instance, the Dice coefficient would calculate the overlap between the candidate GeneRIF and actual GeneRIF. 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 calculated as:

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

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.

Marti Hearst and Presley Nakov developed four derivatives (and Perl code, enhanced by Ravi Teja Bhupatiraju to calculate them) of the classic Dice measurement for the task:

Classic Dice The Dice formula from above applied to words, which were defined as successive alphanumeric characters delimited by white space.

Modified Unigram Dice - This measure gave added weight to terms that occurred multiple times in both strings. In particular, each set of words in a string was a multi-set, with the number of co-occurring words measured by the minimum number of co-occurences.

Bigram Dice - This measure gave additional weight to proper word order. Instead of measuring the unigram Dice coefficient on single words, it measured it on bigrams.

Bigram Phrases - Bigrams do not always represent legitimate phrases. Stop words such as articles and prepositions sometimes occur between content words such that straight bigrams of content words do not represent real phrases. A further measure therefore only included bigrams that did not have intervening stop words filtered.

Official Runs

A total of 14 groups submitted 24 runs. Table 6 lists the runs, sorted by Classic Dice score. The top-ranking run (emc4) came from Erasmus University. The mean and median results are shown at the bottom of the table, followed by the results of a run using simply the document titles.

Most participants found that the GeneRIF text most often came from sentences in the title or abstract of the MEDLINE record, with the title being used most commonly. As such, just using the text of the titles alone achieved a baseline performance that few groups were able to

Table 6 - Official secondary task runs, sorted by classic Dice score.

Run Tag	Classic	Unigram	Bigram	Phrases
emc4	57.83	59.63	46.75	49.11
biotextTask2	53.04	54.65	38.62	41.17
tgIIhugLASt	52.78	54.33	37.72	40.65
UniNEie1	52.28	54.78	37.43	40.35
UniNEie2	51.72	54.27	36.62	39.71
UIowaSecCan	50.68	52.72	35.32	37.87
IBMbtT2	50.47	52.60	34.82	37.91
IUB2003	50.40	52.56	34.83	37.97
NLMUMDLIN	50.36	52.65	35.03	38.34
UniNEie3	49.46	51.42	33.62	36.99
UBGenT2R2	49.40	51.30	33.59	36.99
CSUSMcand	49.31	51.30	34.99	37.80
UBGenT2BL1	49.28	51.25	33.59	36.99
UBGenT2R1	49.03	51.16	33.94	37.35
balscsec1	48.90	50.52	32.36	34.61
we	48.15	49.78	32.31	35.63
nwe	47.62	49.37	31.61	34.80
uwb3	46.48	48.25	29.53	32.82
uwb2	44.41	44.07	2.33	1.80
uwb4	36.28	35.21	22.73	24.52
EDISTFruns2	35.76	35.85	20.05	21.84
tg2hug	35.20	34.57	20.04	21.58
UniNEie4	25.88	25.29	12.03	13.61
UniNEie5	9.42	14.20	0.15	0.17
Mean	45.59	47.16	29.58	32.11
Median	49.30	51.28	33.61	36.99
Titles Only	50.47	52.60	34.82	37.91

outperform. The best approaches (Erasmus (Jelier, Schuemie et al., 2003) and Berkeley (Bhalotia, Nakov et al., 2003)) used classifiers to rank sentences likely to contain the GeneRIF text. No groups much improvement beyond using titles alone.

Future Directions

Despite the limited type of data, relevance judgments, and tasks, the track organizers were pleased with the results and enthusiasm of the participants. We are fortunate to have been awarded a National Science Foundation Information Technology Research grant to provide funding to the track for the next years. The first year's activities also consisted of laying out a roadmap for future iterations of the track. Described in more detail on the track Web site, this will include, over the years, real relevance judgments, use of additional documents beyond MEDLINE, user experiments, and use cases of different types of users.

References

Baxevanis, A. (2003). The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Research,* 31: 1-12.

Bhalotia, G., Nakov, P., et al. (2003). BioText team report for TREC 2003 Genomics Track. *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

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

deBruin, B. and Martin, J. (2003). Finding gene function using LitMiner. *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Jelier, R., Schuemie, M., et al. (2003). Searching for GeneRIFs: concept-based query expansion and Bayes classification. *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Kayaalp, M., Aronson, A., et al. (2003). Methods for accurate retrieval of MEDLINE citations in functional genomics. *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Mitchell, J., Aronson, A., et al. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. *Proceedings of the AMIA 2003 Annual Symposium*, Washington, DC. Hanley & Belfus, 460-464.

O'Neill, G. (2003). Gruber winner Botstein calls for better gene-naming system. Bio-IT World. http://www.bio-itworld.com/news/070903_report2840.html.

Savoy, J., Rasolofo, Y., et al. (2003). Report on the TREC 2003 experiment: Genomics and Web searches. *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Tong, R., Quackenbush, J., et al. (2003). Knowledge-based access to the biomedical literature. *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Wheeler, D., Church, D., et al. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Research,* 31: 28-33.

Yeung, D., Clarke, C., et al. (2003). Task-specific query expansion (MultiText experiments for TREC 2003). *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Zhai, C., Tao, T., et al. (2003). Improving the robustness of language models: UIUC TREC 2003 Genomics and Robust Track experiments. *The Twelfth Text REtrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.