Inf Retrieval (2009) 12:1–15
DOI 10.1007/s10791-008-9076-6

INTRODUCTION

TREC genomics special issue overview

William Hersh · Ellen Voorhees

Received: 10 November 2008 / Accepted: 10 November 2008 / Published online: 1 December 2008

Recent advances in biotechnology have changed the fundamental nature of biological research. Whereas scientists used to be able to manage their modest amount of experimental data in paper notebooks or simple spreadsheets, new tools such as gene chips for measuring gene expression (Mobasheri et al. 2004) or sequence variation (Pennisi 2007) have fundamentally altered their work. Not only do these gene chips generate massive amounts of data (as much as tens of thousands of data points per biological sample), they uncover potential associations and interactions with a wide variety of genes, diseases, and other biological entities. The field devoted to managing, utilizing, and evaluating this data is called *bioinformatics* (Baxevanis and Ouellette 2005), which is sometimes described as the intersection of biology (or biomedicine) and computer science.

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. This requires new approaches to dealing with the biomedical literature, which is the main point of intersection between this field and that of information retrieval (IR) and related disciplines such as text mining.

In the early part of this decade, it became apparent that this situation was ripe for a track at the Text REtrieval Conference (TREC, www.trec.nist.gov), a challenge evaluation for IR organized by the U.S. National Institute of Standards and Technology (NIST, http://www.nist.gov/) (Voorhees and Harman 2005). Started in 1992, TREC has provided a series of challenge evaluations and a forum for presentation of their results. TREC is organized as an annual event at which the tasks are specified and queries and documents are provided to participants. While TREC has historically focused most of its research on textual documents, the field has expanded in recent years with the growth of new information needs (e.g., question-answering, cross-lingual), data types (e.g., sequence data, video) and platforms (e.g., the Web) (Hersh 2003). This special issue is devoted to the TREC Genomics Track, which ran from 2003 to 2007.

W. Hersh (✉)
Oregon Health & Science University, Portland, OR, USA
e-mail: hersh@ohsu.edu

E. Voorhees
National Institute of Standards & Technology, Gaithersburg, MD, USA

The TREC Genomics Track coincides with an increasing amount of biological information resources becoming available in recent years (Galperin 2008). Probably the most important of these are from the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) that maintains most of the NLM's genomics-related databases (Wheeler et al. 2008). As IR has historically focused on text-based data, the NCBI resources of most interest to the IR community include MEDLINE (the bibliographic database of medical literature, accessed by PubMed and other systems) and textbooks such as Online Mendelian Inheritance in Man (OMIM). However, recognizing that literature is often a starting point for data exploration, there is also great interest in resources such as Entrez Gene (Maglott et al. 2007), which serves as a switchboard to integrate gene information as well as provide annotation of its function using the widely accepted GeneOntology (GO) (Anonymous 2008). PubMed also provides linkages to full-text journal articles on the Web sites of publishers. Additional genomics resources exist beyond the NCBI, such as the model organism genome databases (Bahls et al. 2003). As with the NCBI resources, these resources provide rich linkage and annotation.

Both the IR and bioinformatics communities have long histories of forums for evaluation of methods. The latter has the well-known Critical Assessment of Methods of Protein Structure Prediction (CASP) initiative for protein structure prediction (Moult et al. 2003; Venclovas et al. 2003). More recently, challenge evaluations have been initiated for researchers interested in information extraction (IE) (Hirschman et al. 2002), including the Knowledge Discovery from Databases (KDD) Cup (Yeh et al. 2003) and the BioCreative initiative (Hirschman et al. 2005).

With the exception of the Genomics Track, TREC has not focused on biomedical content. The TREC Genomics Track (http://ir.ohsu.edu/genomics/) was one of the largest and longest running challenge evaluations in biomedicine. The tasks of each year are listed in Table 1 and described in more detail in this paper. Instructions for obtaining the test collections for research use are available on the track's Web site. The remainder of this paper will describe the details of the specific tasks of the TREC Genomics Track as well as the papers that were accepted for inclusion in this special issue.

1 Ad hoc retrieval tasks

A major focus early in the TREC Genomics Track was on ad hoc retrieval. In 2003, before funding for substantial relevance judgments was available, a 1-year subset of MEDLINE was obtained and topics for the test collection were gene names. Documents were designated as relevant if a Gene Reference into Function (GeneRIF) (Mitchell et al. 2003) was available for the article (Hersh and Bhupatiraju 2003). GeneRIFs are a textual annotation about the function of the gene that is maintained by NCBI along with MEDLINE and other databases. This clearly underestimated the number of relevant documents, since at the time there was a modest number of GeneRIFs annotated, probably not allowing true reflection of system performance. But it did allow researchers to start working with biologically oriented documents and topics.

When funding became available in 2004, the ad hoc retrieval task was expanded, allowing a larger test collection with true relevance judgments to be developed. This task modeled the situation of a genomics researcher entering a new area and having an information need in that area using an IR system to access the biomedical scientific literature. The document collection was based on a ten-year subset of MEDLINE. The rationale for using MEDLINE was that despite being in an era of readily available full-text journals

Table 1 Tasks of the TREC genomics track (http://ir.ohsu.edu/genomics/)

Year	Task description	Document collection	Topics
2003 (Hersh and Bhupatiraju 2003)	Ad hoc retrieval	A 1-year (4/2002–4/2003) subset of 525,938 MEDLINE records	Gene names, with the goal of finding all MEDLINE references that focus on the basic biology of the gene or its protein products from the designated organism
2003 (Hersh and Bhupatiraju 2003)	Annotation of gene reference into functions (GeneRIFs) (Mitchell et al. 2003) from article titles and abstracts	139 articles that had been assigned GeneRIFs, derived from all articles appearing in five journals during the latter half of 2002	Assigned GeneRIFs
2004 (Hersh et al. 2004, 2006a)	Ad hoc retrieval	A ten-year subset (1994–2003) of 4,591,008 MEDLINE records	50 information needs statements with title, information need, and context (background)
2004 (Hersh et al. 2004; Cohen and Hersh 2006)	Categorization of documents containing data about gene function suitable for "triage" to annotators assigning Gene Ontology (GO) codes for Mouse Genome Informatics (MGI) database	A 3-year set of 11,880 full-text articles for three journals obtained from Highwire Press	N/A
2005 (Hersh et al. 2005)	Ad hoc retrieval	A 10-year subset (1994–2003) of 4,591,008 MEDLINE records	50 information needs statements similar to 2004 but classified into one of five Generic Topic Types (GTTs)
2005 (Hersh et al. 2005)	Categorization of documents containing data about gene function suitable for "triage" to annotators assigning GO codes or identifying for inclusion into databases about tumor biology, embryologic gene expression, or alleles of mutant phenotypes for MGI	A 3-year set of 11,880 full-text articles for three journals obtained from Highwire Press	N/A
2006 (Hersh et al. 2006b)	Passage retrieval (from part of sentence to paragraph in length) with linkage to five entities (e.g., genes, proteins) and the source article	Collection of 162,259 full-text HTML documents from 49 journals that publish electronically via Highwire Press	28 question statements based on GTTs
2007 (Hersh et al. 2007)	Entity-based question-answering based on retrieval of passages linked to 14 entities and the source article	Collection of 162,259 full-text HTML documents from 49 journals that publish electronically via Highwire Press	36 question statements based on the 14 entities

(usually requiring a subscription), many users still entered the biomedical literature through searching MEDLINE. As such, there were still strong motivations to improve the effectiveness of searching MEDLINE. The ad hoc retrieval task ran in the TREC 2004 (Hersh et al. 2004, 2006a) and TREC 2005 (Hersh et al. 2005) Genomics Tracks.

The MEDLINE subset consisted of 10 years of completed citations from the database inclusive from 1994 to 2003. Records were extracted using the Date Completed (DCOM) field for all references in the range of 19940101–20031231. This provided a total of 4,591,008 records, which was about one-third of the full MEDLINE database. The data included all of the PubMed fields identified in the MEDLINE Baseline record. The subset was provided in the "MEDLINE" format, consisting of ASCII text with fields indicated and delimited by two to four character abbreviations. The size of the file uncompressed was about 9.5 GB. In this subset, there were 1,209,243 (26.3%) records without abstracts.

Topics for the ad hoc retrieval task were based on information needs collected from real biologists. In the 2004 track, simple information needs were collected and formatted into 50 topics with the following fields:

- ID—identifier
- Title—abbreviated statement of information need
- Information need—full statement information need
- Context—background information to place information need in context

In the 2005 track, instead of soliciting free-form biomedical questions, a set of five generic topic templates (GTTs) derived from an analysis of the topics from the 2004 track and other known biologist information needs were developed (see Table 2). These GTTs consisted of semantic types, such as genes or diseases, placed in the context of commonly queried biomedical questions. After development of the GTTs, biologists were interviewed to obtain specific information needs that conformed to each GTT. The topics did not have to fit precisely into the GTTs, but had to come close, i.e., have all the required semantic types. Ten information needs for each GTT were selected for inclusion in the 2005 track to obtain 50 topics.

Relevance judgments for both years were performed carrying out the usual pooling method of TREC, where the top-ranking results of all official runs submitted by track participants were pooled. The relevance judges in general were individuals who had backgrounds in either biology or medicine. The relevance assessors judged each document

Table 2 Generic topic types and example sample topics for the TREC 2005 genomics track, with the semantic types in each generic topic type (GTT) underlined (Hersh et al. 2005)

Generic topic type	Example sample topic
Find articles describing standard methods or protocols for doing some sort of experiment or procedure	Method or protocol: GST fusion protein expression in Sf9 insect cells
Find articles describing the role of a gene involved in a given disease	Gene: DRD4 Disease: Alcoholism
Find articles describing the role of a gene in a specific biological process	Gene: Insulin receptor gene Biological process: Signaling tumorigenesis
Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease	Genes: HMG and HMGB1 Disease: Hepatitis
Find articles describing one or more mutations of a given gene and its biological impact	Gene with mutation: Ret Biological impact: Thyroid function

for the specific topic as definitely relevant (DR), possibly relevant (PR), or not relevant (NR). For the official results, which required binary relevance judgments, documents that were rated DR or PR were considered relevant. In the 2005 track, articles had to describe a specific gene, disease, impact, mutation, etc. and not just the concept in general. In addition, relevance judges were given more explicit instructions relative to the GTTs:

- Relevant article must describe how to conduct, adjust, or improve a standard, a, new method, or a protocol for doing some sort of experiment or procedure.
- Relevant article must describe some specific role of the gene in the stated disease or biological process.
- Relevant article must describe a specific interaction (e.g., promote, suppress, inhibit, etc.) between two or more genes in the stated function of the organ or the disease.
- Relevant article must describe a mutation of the stated gene and the particular biological impact(s) that the mutation has been found to have.

For both the 2004 and 2005 tracks, the primary measure of performance was mean average precision (MAP) (Buckley and Voorhees 2005). Research groups were also required to classify their runs into one of three categories:

- Automatic—no manual intervention in building queries
- Manual—manual construction of queries but no further human interaction
- Interactive—completely interactive construction of queries and further interaction with system output

In the 2004 track, the best results were obtained by a combination of Okapi weighting (BM25 for term frequency but with standard inverse document frequency), Porter stemming, expansion of symbols by LocusLink and MeSH records, query expansion, and use of all three fields of the topic (title, need, and context) (Fujita 2004). These achieved a MAP of 0.4075. When the language modeling technique of Dirichlet-Prior smoothing was added, an even higher MAP of 0.4264 was obtained. Another group achieved high-ranking results with a combination of approaches that included Okapi weighting, query expansion, and various forms of domain-specific query expansion (including expansion of lexical variants as well as acronym, gene, and protein name synonyms) (Buttcher et al. 2004). Approaches that attempted to map to controlled vocabulary terms did not fare as well (Aronson et al. 2004; Nakov et al. 2004; Seki et al. 2004). As always in TREC, many groups tried a variety of approaches, beneficial or otherwise, but usually without comparing common baseline or running exhaustive experiments, making it difficult to discern exactly what techniques provided benefit and which techniques could be productively combined or were essentially equivalent.

Somewhat similar results were obtained in the 2005 track. As with 2004, the basic Okapi with good parameters gave good baseline performance for a number of groups. Manual synonym expansion of queries gave the highest MAP of 0.302 (Huang et al. 2005), although automated query expansion did not fare as well (Ando et al. 2005; Aronson et al. 2005). Relevance feedback was found to be beneficial, but worked best without term expansion (Zheng et al. 2005).

Follow-up research with the TREC Genomics Track ad hoc retrieval test collections has yielded a variety of findings. One study assessed word tokenization, stemming, and stop word removal, finding that varying strategies for the first resulted in substantial performance impact while changes in the latter two had minimal impact. Tokenization of genomics text can be challenging due to the use of a wide variety of symbols, including numbers, hyphens, super- and sub-scripts, and characters in non-English languages (e.g.,

Greek) (Jiang and Zhai 2007). Another study found value for language modeling approaches to term weighting. Other studies have assessed improving the related-articles feature of PubMed (Lin and Wilbur 2007) and categorizing articles containing data for inclusion in comparative effectiveness reviews of drug efficacy (Cohen et al. 2006).

2 Summarization tasks

Another task run in 2003 was a summarization task, where researchers where challenged with nominating the annotation text of the GeneRIF (Hersh and Bhupatiraju 2003). This was akin to text summarization, where systems had to nominate an excerpt of text that summarized what was in the document, which in this case was the full text of 139 documents for which GeneRIFs were available. Performance was measured between overlap with the nominated text and the GeneRIF annotation using the Dice coefficient. The best-performing systems were found to use the text of the title of the article, achieve Dice coefficient scores approaching 60%.

3 Text categorization tasks

A second task in 2004 and 2005 was a biomedical text categorization task. The mail goal of the task was to "triage" articles as requiring further analysis for human annotators in the Mouse Genome Informatics (MGI) system (http://www.informatics.jax.org/). Systems were required to classify full-text documents from a 2-year span (2002–2003) of three journals, with the first year's (2002) documents comprising the training data and the second year's (2003) documents making up the test data.

One of the goals of MGI is to provide structured, coded annotation of gene function from the biological literature. Human curators identify genes and assign Gene Ontology (GO) and other codes about gene function with another code describing the type of experimental evidence supporting assignment of the code. The huge amount of literature requiring curation creates a challenge for MGI, as their human resources are not unlimited. As such, they employ a three-step process to identify the papers most likely to describe gene function:

1. About mouse—The first step is to identify articles about mouse genomics biology. The full text of articles from several hundred journals are searched for the words *mouse*, *mice*, or *murine*. Articles passing this step are further analyzed for inclusion in MGI. At present, articles are searched in a Web browser one at a time because full-text searching is not available for all of the journals included in MGI.
2. Triage—The second step is to determine whether the identified articles should be sent for curation. MGI curates articles not only for GO terms, but also for other aspects of biology, such as gene mapping, gene expression data, phenotype description, and more. The goal of this triage process is to limit the number of articles sent to human curators for more exhaustive analysis. Articles that pass this step go into the MGI system with a tag for GO, mapping, expression, etc. The rest of the articles do not go into MGI.
3. Annotation—The third step is the actual curation with GO and other terms. In the case of GO codes, curators identify genes for which there is experimental evidence to warrant assignment of codes, with another code for each indicating the type of experimental evidence. There can be more than one gene assigned one GO code in a given paper and there can be more than one GO code assigned for a gene (i.e., potentially a many-to-many relationship).

Table 3 Best and median utility scores for each subtask of the TREC Genomics text categorization task, adapted from Hersh et al. (2005)

Subtask	Best utility	Median utility
A (alelle)	0.871	0.7773
E (expression)	0.8711	0.6413
G (GO annotation)	0.587	0.4575
T (tumor)	0.9433	0.761

The TREC Genomics text categorization tasks focused on triage of articles since this function was believed by MGI to have the most value in automating. In addition, challenge evaluations such as Biocreative (described above) were already investigating annotation. The triage task required a system to decide whether an article should be sent to a curator for annotation. Performance was assessed by the utility measure from the TREC Filtering Track (http://trec.nist.gov/data/filtering.html), with the parameters u_r and u_{nr} tuned for each specific triage subtask. In TREC 2004, the triage task was to assign articles for GO annotation, whereas in 2005, the task was expanded to include triage for inclusion in databases about tumor biology (Krupke et al. 2005), embryologic gene expression (Hill et al. 2004), and alleles of mutant phenotypes (Strivens and Eppig 2004).

The documents for the categorization task consisted of articles from three journals over 2 years published by Highwire Press (http://www.highwire.org/). The journals available and used by the task were *Journal of Biological Chemistry* (JBC), *Journal of Cell Biology* (JCB), and *Proceedings of the National Academy of Science* (PNAS). Each of the papers from these journals was provided in SGML format based on Highwire's Document Type Definition (DTD). Articles from the year 2002 were assigned as training data and articles from 2003 were assigned as test data.

The results from different groups are summarized in Table 3 and papers describing the task (Hersh et al. 2005; Cohen and Hersh 2006). These groups used a variety of NLP and machine learning tasks, with a wide range of results. One notable finding across all groups was the GO triage subtask was substantially more difficult than the tumor biology, embryologic gene expression, or alleles of mutant phenotypes subtasks. Over the two years that this task was repeated, very little could be done to improve triage of articles for GO annotation beyond the presence of the MeSH term Mice. However, performance on the other three subtasks was generally very good. Some additional work has used a subset of the TREC Categorization data to assess the detection of figures and their types for use as features (Shatkay et al. 2006).

4 Question–answering tasks

In the latter 2 years of the track, the focus shifted to question–answering in the biomedical domain. In 2006 and 2007, the track implemented a task that covered *entity-based question answering* (Hersh et al. 2006b, 2007). The rationale for the task was that information seekers, especially users of the biomedical literature, frequently desire something between strictly defined IR and IE, i.e., a system that provides short, specific answers to questions and that puts the answers in context by providing supporting information and linking to original sources. As such, the track developed a new task that focused on retrieval of short passages (from phrase to sentence to paragraph in length) that specifically addressed an information need, along with linkage to the location in the original source document.

Topics were expressed as questions and systems were measured on how well they retrieved relevant information at the passage, aspect, and document levels. Systems were

required to return passages linked to source documents, while relevance judges not only rated the passages, but also grouped them by aspect. For this task, aspect was defined similar to its definition in the TREC Interactive Track aspectual recall task (Hersh 2001), representing answers that covered a similar portion of a full answer to the topic question. The track also drew upon experience in passage retrieval from the previous TREC High Accuracy Retrieval from Documents (HARD) Track (Allan 2003, 2004).

The documents for this task came from a new full-text biomedical corpus, as track members had also advocated a move from bibliographic (MEDLINE) to full-text documents (journal articles). Permission was obtained from a number of publishers who used Highwire Press (http://www.highwire.org/) for electronic distribution of their journals. Those publishers agreed to allow use of their full text in HTML format, which preserved formatting, structure, table and figure legends, etc. The document collection was derived from 49 journals and contained 162,259 documents, which was about 12.3 GB in size when uncompressed. In addition to the full-text data, the NLM provided MEDLINE records for the full-text documents in the collection.

Some additional files were made available:

- A text file, metadata.txt, listed the original URL of the article, the file name in this collection, and its size in kilobytes. The name of each document file was its Pubmed Identifier (PMID) plus the extension .html, which facilitated accessing the associated MEDLINE record.
- Another file, legalspans.txt, contained all "legal spans" for all documents in the collection. Legal spans were defined as any contiguous text >0 characters in length not including any HTML paragraph tags, defined as any tag that started with <P or </P (case insensitive). There were a total of 12,641,127 legal spans in the collection. These were used to define allowed passages in the pooling and evaluation process, and to limit the size of the passages that needed reviewing by the expert judges.

Retrieved passages could contain any span of text that did not include any part of an HTML paragraph tag (i.e., one starting with <P or </P). Because there was some confusion about the different types of passages, the following terms were defined:

- Nominated passages—These were the passages that systems nominated in their runs and were scored in the passage retrieval evaluation. To be legal, these passages had to be a subset of a maximum-length legal span.
- Maximum-length legal spans—These were all the passages obtained by delimiting the text of each document by the HTML paragraph tags. As noted below, nominated passages could not cross an HTML paragraph boundary. So these spans represented the longest possible passage that could be designated as relevant. These spans were also used to build pools for the relevance judges. The judges did not need to designate the entire span as relevant, and could select just a part of the span as the relevant passage.
- Relevant passages—These were the spans that the judges designated as definitely or possibly relevant.

The first running of the task took place in 2006, with the topics expressed as questions (Hersh et al. 2006b). They were derived from the set of biologically relevant questions based on the GTTs developed for the 2005 track (Hersh et al. 2005). The questions (and GTTs) all had the general format of containing one or more biological objects and processes and some explicit relationship between them. The biological objects might be genes, proteins, gene mutations, etc. The biological process could be physiological processes or

diseases. The relationships could be anything, but were typically verbs such as *causes*, *contributes to, affects, associated with*, or *regulates*.

The relevance assessments were done by the usual TREC method of pooling the top-ranking passages from different groups that submitted official runs. For each topic, a pool of passages was created that consisted of maximum-length spans from those passages there were retrieved. The relevance judges were experts (usually having a PhD in biology or a related life science) who were provided with guidelines and a training session to improve the judging process. To assess relevance, judges were instructed to break down the question into required elements (e.g., the biological entities and processes that make up the GTT) and isolate the minimum contiguous substring that answered the question. In general, a passage was definitely relevant if it contained all required elements of the question and it answered the question. A passage was possibly relevant if it contained the majority of required elements, missing elements were within the realm of possibility (i.e. more general terms are mentioned that probably include the missing elements), and it possibly answered the question.

After determining the "best" answer passages, judges were instructed to group them into related concepts and then assign one or more Medical Subject Headings (MeSH) terms (possibly with subheadings) to capture similarities and differences among retrieved passage aspects. They were told to use the most specific MeSH term, with the option of adding subheadings, similar to the NLM literature indexing process. If one term was insufficient to denote all aspects of the gold standard passage, judges assigned additional MeSH terms. All passages judged as definitely or possibly relevant were required to have a gold standard passage and at least one MeSH term. For all the topics, the mean number of relevant passages was 35 (range 3–593), with a mean relevant passage length of 400 characters (range 27–6928). There were an average of 22 distinct relevant aspects per topic (range 7–96).

For this entity-based, question-answering task, there were three levels of retrieval performance measured: passage retrieval, aspect retrieval, and document retrieval. Each of these provided insight into the overall performance for a user trying to answer the given topic questions. Each was measured by some variant of MAP.

- Passage-level MAP—This measure used a variation of MAP, computing individual precision scores for passages based on character-level precision, using a variant of a similar approach used for the TREC 2004 HARD Track (Allan 2004). For each nominated passage, the number of characters that overlapped with those deemed relevant by the judges in the gold standard were determined. For each relevant retrieved passage, precision was computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point. Similar to regular MAP, remaining relevant passages that were not retrieved (no overlap with any nominated passages) were added into the calculation as well, with precision set to 0 for these relevant non-retrieved gold standard passages. Then the mean of these average precisions over all topics was calculated to compute the MAP for passages.
- Aspect-level MAP—Aspect retrieval was measured using the average precision for the aspects of a topic, averaged across all topics. To compute this, for each submitted run, the ranked passages were transformed to two types of values, either the aspect(s) of the gold standard passage that the submitted passage overlapped with or the value "not relevant". This resulted was a ranked list, for each run and each topic, of lists of aspects per passage, Nonrelevant passages had empty lists of aspects. Because of the uncertainty of the value for a user of a repeated aspect (e.g., same aspect occurring

again further down the list), these were discarded from the output to be analyzed. For the remaining aspects of a topic, precision for the retrieval of each aspect was computed as the fraction of relevant passages for the retrieved passages up to the first passage with the aspect under consideration. These fractions at each point of first aspect retrieval were then averaged together to compute the average aspect precision. Taking the mean over all topics produced the final aspect-based MAP.

- Document-level MAP—For the purposes of this measure, any PMID that had a passage associated with a topic ID in the set of gold standard passages was considered a relevant document for that topic. All other documents were considered not relevant for that topic. System run outputs were collapsed by PMID document identifier, with the documents appearing in the same order as the first time the corresponding PMID appeared in the nominated passages for that topic. For a given system run, average precision was measured at each point of correct (relevant) recall for a topic. The MAP was the mean of the average precisions across topics.

As shown in Table 4, document MAP scores were highest, followed by aspect, and then passage, although these scores were not directly comparable since they measured precision at recall of different things. There was a general, though far from perfect, correlation between the measures across all submissions. It was clear from the results and techniques of the top-performing groups in passage retrieval that certain approaches were quite effective. In particular, "trimming" passages to shorten them was done in all the runs with the highest passage MAP. Indeed, because non-content manipulations of passages had substantial effects on passage MAP, an alternative passage MAP (PASSAGE2) that calculated MAP as if each character in each passage were a ranked document was developed for additional analysis and used in the TREC 2007 Genomics Track.

A further analysis showed that four system factors were associated with the best performance in passage MAP (Rekapalli et al. 2007):

- Normalization of keywords in the query into root forms
- Non-use of the Entrez Gene thesaurus for synonym terms expansion
- Unit of text retrieved using respective IR algorithms at sentence level
- Passage "trimming" to best sentence

The TREC 2007 Genomics Track continued with the same task and document collection, but some modifications to the topics and relevance judging were made, along with adoption

Table 4 Overall results from TREC 2006–2007 genomics track task, adapted from Hersh et al. (2006b, 2007)

	Passage2 MAP	Passage MAP	Aspect MAP	Document MAP
TREC 2006				
Min	0.0007	0.0019	0.0110	0.0198
Median	0.0345	0.0316	0.1581	0.3083
Mean	0.0392	0.0347	0.1643	0.2887
Max	0.1486	0.1012	0.4411	0.5439
TREC 2007				
Min	0.0008	0.0029	0.0197	0.0329
Median	0.0377	0.0565	0.1311	0.1897
Mean	0.0398	0.0560	0.1326	0.1862
Max	0.1148	0.0976	0.2631	0.3286

of a new official measure of passage retrieval performance (PASSAGE2) (Hersh et al. 2007). There were 36 official topics for the track in 2007, which were in the form of questions asking for lists of specific entities. As in the past, information needs were gathered from working biologists. In addition to asking about information needs, there biologists were asked if their desired answer was a list of a certain type of entity, such as genes, proteins, diseases, mutations, etc., and if so, to designate that entity type. An example topic was:

What [GENES] are genetically linked to alcoholism?

Answers to this question were passages that related one or more entities of type GENE to alcoholism. For example, a valid and relevant answer to this topic would be, *The DRD4 VNTR polymorphism moderates craving after alcohol consumption* (from PMID 11950104). And the GENE entity supported by this statement would be DRD4. Table 5 shows the entities, their definitions, potential sources of terms, and topics with each entity type.

Table 5 TREC 2007 genomics track entities, their definitions, potential sources of terms, and topics with each entity type

Entity type	Definition	Topics with entity type
Antibodies	Immunoglobulin molecules having a specific amino acid sequence by virtue of which they interact only with the antigen (or a very similar shape) that induced their synthesis in cells of the lymphoid series (especially plasma cells)	1
Biological substances	Chemical compounds that are produced by a living organism	3
Cell or tissue types	A distinct morphological or functional form of cell, or the name of a collection of interconnected cells that perform a similar function within an organism	2
Diseases	A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown	1
Drugs	A pharmaceutical preparation intended for human or veterinary use	2
Genes	Specific sequences of nucleotides along a molecule of DNA (or, in the case of some viruses, RNA) which represent functional units of heredity	11
Molecular functions	Elemental activities, such as catalysis or binding, describing the actions of a gene product or bioactive substance at the molecular level	2
Mutations	Any detectable and heritable change in the genetic material that causes a change in the genotype and which is transmitted to daughter cells and to succeeding generations	1
Pathways	A series of biochemical reactions occurring within a cell to modify a chemical substance or transduce an extracellular signal	2
Proteins	Linear polypeptides that are synthesized on ribosomes and may be further modified, crosslinked, cleaved, or assembled into complex proteins with several subunits	5
Strains	A genetic subtype or variant of a virus or bacterium	2
Signs or symptoms	A sensation or subjective change in health function experienced by a patient, or an objective indication of some medical fact or quality that is detected by a physician during a physical examination of a patient	1
Toxicities	A measure of the degree and the manner in which which something is toxic or poisonous to a living organism	2
Tumor types	An abnormal growth of tissue, originating from a specific tissue of origin or cell type, and having defined characteristic properties, such as a recognized histology	1

Relevance judging was once again by pooling of top-ranking passages retrieved by participating groups. Judges were required to have significant domain knowledge, typically in the form of a PhD in a life science. They were trained using a 12-page manual and a one-hour videoconference. They were given the following instructions:

1. Review the topic question and identify key concepts.
2. Identify relevant paragraphs and select minimum complete and correct excerpts.
3. Develop a topic-specific controlled vocabulary for entities based on the relevant passages and contained entities and code entities for each relevant passage based on this vocabulary.

There were an average of 124.8 relevant passages containing an average of 72.3 aspects from 69.2 relevant documents per topic. The mean relevant passage length was 968, with an average of 1.63 aspects per relevant passage.

As in all other years of the track, there were a variety of approaches used in 2007 that demonstrated varying levels of benefit. In the track overview paper, we tried to cluster runs by the features they employed and then compare results with the different measures. In the 2007 track, we found clusters of approaches that included query expansion, use of language models, and varying units of initial passage retrieval. The results made it clear that the explicit methods used within each of these was more important than the general approach. Two approaches found to achieve benefit in many but not all instances were query expansion with synonyms and retrieval based on paragraph-sized or larger units (Hersh et al. 2007).

Another recurring aspect of interpreting results in the 2007 track was the difficulty in interpreting results over all groups due to inadequate reporting in proceedings papers, use of variable baselines for comparison, and insufficiently exhaustive experimentation. These findings led to the call for papers for this special issue advocating sufficiently comprehensive experimentation.

5 Special issue papers

Nine papers were submitted for review for this issue, of which four were accepted. Three of these papers explore aspects of query expansion while the fourth looks at factors related to the quality of the relevance judging process. The focus on query expansion is appropriate, given that this method was among the most prominent in improving retrieval performance over the years but also had substantial variation in what particular approaches did and did not work well. Some of these papers also address other areas shown to be effective but variable, such as passage retrieval size.

The paper by Stokes et al. focuses on successful factors for query expansion using the 2006 track documents and topics (Stokes et al. 2008). Through a combination of approaches, they are able to advance passage MAP by 185% over the basic Okapi retrieval system. Through exhaustive experimentation, they find several factors most highly associated with success. One of these is the re-ranking of concepts that give the most weight to occurrence or not as opposed to the frequency of occurrence. It is also found that normalization of expansion terms is important. They also find the most benefit of term expansion comes from formal instead of ad hoc (e.g., co-occurrence or hierarchy based) terminologies and in particular from gene (as opposed to general biomedical term or abbreviation) synonyms.

Lu et al. focus on gene synonym expansion only, omitting experimentation on other aspects of retrieval, such as tokenization or stemming (Lu et al. 2008a). They find with

gene-only topics (i.e., the 2003 collection), query name expansion helps substantially. With the more verbose queries of the other years' data (2004–2007), normalizing to gene names is crucial [similar to Stokes et al. (2008)] and performance can be improved with a variety of language model approaches.

Lu et al. also focus on gene synonym expansion, omitting topics that do not allow such expansion (Lu et al. 2008b). Using documents and topics from the 2006 and 2007 tracks, they find modest improvement with expansion based on the automated term mapping process in the operational PubMed system. However, of practical concern, they note that the benefit yields minimal improvements in the portion of output users would be likely to view, i.e., the top 20–30 documents of the search. Another interesting finding in this study warranting further research is the better MAP obtained by ranking using classic term weighting schemes over the reverse chronological sorting used as a default by PubMed.

The final paper in this issue looks at the relevance judging process, which was led from 2005–2007 by the paper's first author (Roberts et al. 2008). There are many lessons learned, as this paper shows, in that the process works best when judges are given explicit training and instructions and have domain expertise. One unfortunate consequence noted by the authors was the changing of the task each year, which did not enable them to assess the benefits of specific improvements in the process.

6 Conclusions

The TREC Genomics Track for the most part achieved what it set out to accomplish, which was to provide test collections for experimentation in IR in the genomics domain. The different collections, based on varying tasks, showed that some approaches appeared to benefit retrieval and related tasks specifically in this domain. As with all TREC activity, the short cycle of experimentation and reporting of results has prevented more detailed investigation of different approaches. However, there emerged some evidence that some resources from the genomics/bioinformatics could contribute to improving retrieval, especially controlled lists of terminology used in query expansion, although their improvement over standard state-of-the-art IR was not substantial.

The existence and continued availability of the collections will hopefully encourage researchers to delve into other aspects of IR in the genomics and larger biomedical domain. There are still many avenues of experimentation that may yield improvements in retrieval performance, which in turn will potentially improve scientific discovery in genomics and provide benefits to larger human health. The TREC Genomics Track web site will be maintained, including instructions for accessing the data collections for research purposes.

Acknowledgements The TREC Genomics Track was supported by NSF Grant ITR-0325160. The track also appreciated the help of Lori Buckland and others at NIST for help in its administration. We were also very grateful to the National Library of Medicine, Highwire Press, and the Mouse Genome Informatics Project for providing data for use in the track.

References

Allan, J. (2003). HARD Track overview in TREC 2003—high accuracy retrieval from documents. *The Twelfth Text REtrieval Conference-TREC 2003* (pp. 24–37). Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec12/papers/HARD.OVERVIEW.pdf.

Allan, J. (2004). HARD Track overview in TREC 2004—high accuracy retrieval from documents. *The Thirteenth Text REtrieval Conference* (*TREC 2004*). Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/HARD.OVERVIEW.pdf.

Ando, R., Dredze, M., et al. (2005). TREC 2005 genomics track experiments at IBM Watson. *The Fourteenth Text Retrieval Conference Proceedings* (*TREC 2005*). Gaithersburg, MD: National Institute for Standards and Technology. http://trec.nist.gov/pubs/trec14/papers/ibm-tjwatson.geo.pdf.

Anonymous. (2008). The gene ontology project in 2008. *Nucleic Acids Research, 36*, D440–D444.

Aronson, A., Demmer, D., et al. (2004). Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations. *The Thirteenth Text Retrieval Conference* (*TREC 2004*). Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/nlm-umd-ul.geo.pdf.

Aronson, A., Demner-Fushman, D., et al. (2005). Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. *The Fourteenth Text REtrieval Conference Proceedings* (*TREC 2005*). Gaithersburg, MD: National Institute for Standards and Technology. http://trec.nist.gov/pubs/trec14/papers/nlm-umd.geo.pdf.

Bahls, C., Weitzman, J., et al. (2003). Biology's models. *Scientist* (*Philadelphia, PA*), *5*. http://www.the-scientist.com/yr2003/jun/feature_030602.html.

Baxevanis, A., & Ouellette, B. (2005). *Bioinformatics: A practical guide to the analysis of genes and proteins* (3rd ed.). Hoboken, NJ: Wiley.

Buckley, C., & Voorhees, E. (2005). Retrieval system evaluation. In E. Voorhees & D. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 53–75). Cambridge, MA: MIT Press.

Buttcher, S., Clarke, C., et al. (2004). Domain-specific synonym expansion and validation for biomedical information retrieval (MultiText experiments for TREC 2004). *The Thirteenth Text Retrieval Conference* (*TREC 2004*). Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uwaterloo-clarke.geo.pdf.

Cohen, A., & Hersh, W. (2006). The TREC 2004 genomics track categorization task: Classifying full-text biomedical documents. *Journal of Biomedical Discovery and Collaboration, 1,* 4. http://www.j-biomed-discovery.com/content/1/1/4. doi:10.1186/1747-5333-1-4.

Cohen, A., Hersh, W., et al. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association, 13*, 206–219. doi:10.1197/jamia.M1929.

Fujita, S. (2004). Revisiting again document length hypotheses—TREC 2004 genomics track experiments at Patolis. *The Thirteenth Text Retrieval Conference* (*TREC 2004*). Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/patolis.geo.pdf.

Galperin, M. (2008). The molecular biology database collection: 2008 update. *Nucleic Acids Research, 36*, D2–D4. doi:10.1093/nar/gkm1037.

Hersh, W. (2001). Interactivity at the text retrieval conference (TREC). *Information Processing and Management, 37*, 365–366. doi:10.1016/S0306-4573(00)00052-2.

Hersh, W. (2003). *Information retrieval: A health and biomedical perspective* (2nd ed.). New York: Springer-Verlag. http://www.irbook.info.

Hersh, W., & Bhupatiraju, R. (2003). TREC genomics track overview. *The Twelfth Text Retrieval Conference* (*TREC 2003*) (pp. 14–23). Gaithersburg, MD: NIST. http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.

Hersh, W., Bhuptiraju, R., et al. (2004). TREC 2004 genomics track overview. *The Thirteenth Text Retrieval Conference* (*TREC 2004*). Gaithersburg, MD: National Institute for Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf.

Hersh, W., Cohen, A., et al. (2005). TREC 2005 genomics track overview. *The Fourteenth Text Retrieval Conference* (*TREC 2005*). Gaithersburg, MD: National Institute for Standards and Technology. http://trec.nist.gov/pubs/trec14/papers/GEO.OVERVIEW.pdf.

Hersh, W., Bhupatiraju, R., et al. (2006a). Enhancing access to the bibliome: the TREC 2004 genomics track. *Journal of Biomedical Discovery and Collaboration, 1*, 3. http://www.j-biomed-discovery.com/content/1/1/3. doi:10.1186/1747-5333-1-3.

Hersh, W., Cohen, A., et al. (2006b). TREC 2006 genomics track overview. *The Fifteenth Text Retrieval Conference* (*TREC 2006*) (pp. 52–78). Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec15/papers/GEO.OVERVIEW.pdf.

Hersh, W., Cohen, A., et al. (2007). TREC 2007 genomics track overview. *The Sixteenth Text Retrieval Conference* (*TREC 2007*). Gaithersburg, MD: National Institute for Standards and Technology. http://ir.ohsu.edu/genomics/trec-07-genomics.pdf.

Hill, D., Begley, D., et al. (2004). The mouse gene expression database (GXD): Updates and enhancements. *Nucleic Acids Research, 32*, D568–D571. doi:10.1093/nar/gkh069.

Hirschman, L., Park, J., et al. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics (Oxford, England), 18*, 1553–1561. doi:10.1093/bioinformatics/18.12.1553.

Hirschman, L., Yeh, A., et al. (2005). Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics, 6*, S1. http://www.biomedcentral.com/1471-2105/6/S1/S1. doi:10.1186/1471-2105-6-S1-S1.

Huang, X., Zhong, M., et al. (2005). York University at TREC 2005: Genomics track. *The Fourteenth Text REtrieval Conference Proceedings* (*TREC 2005*). Gaithersburg, MD: National Institute for Standards and Technology. http://trec.nist.gov/pubs/trec14/papers/yorku-huang2.geo.pdf.

Jiang, J., & Zhai, C. (2007). An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval, 10*, 341–363. doi:10.1007/s10791-007-9027-7.

Krupke, D., Naf, D., et al. (2005). The mouse tumor biology database: Integrated access to mouse cancer biology data. *Experimental Lung Research, 31*, 259–270. doi:10.1080/01902140490495633.

Lin, J., & Wilbur, W. (2007). PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics, 8*, 423. http://www.biomedcentral.com/1471-2105/8/423. doi:10.1186/1471-2105-8-423.

Lu, Y., Fang, H., et al. (2008a). An empirical study of gene synonym query expansion in biomedical information retrieval. *Information Retrieval*. doi:10.1007/s10791-008-9075-7.

Lu, Z., Kim, W., et al. (2008b). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*. doi:10.1007/s10791-008-9074-8.

Maglott, D., Ostell, J., et al. (2007). Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research, 35*, D26–D31. doi:10.1093/nar/gkl993.

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

Mobasheri, A., Airley, R., et al. (2004). Post-genomic applications of tissue microarrays: Basic research, prognostic oncology, clinical genomics and drug discovery. *Histology and Histopathology, 19*, 325–335.

Moult, J., Fidelis, K., et al. (2003). Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins, 53*, 334–339. doi:10.1002/prot.10556.

Nakov, P., Schwartz, A., et al. (2004). BioText team experiments for the TREC 2004 genomics track. *The Thirteenth Text Retrieval Conference* (*TREC 2004*). Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/ucal-berkeley.geo.pdf.

Pennisi, E. (2007). Breakthrough of the year—human genetic variation. *Science, 318*, 1842–1843. doi:10.1126/science.318.5858.1842.

Rekapalli, H., Cohen, A., et al. (2007). A comparative analysis of retrieval features used in the TREC 2006 Genomics Track passage retrieval task. *Proceedings of the AMIA 2007 Annual Symposium*. Chicago, IL: American Medical Informatics Association.

Roberts, P., Cohen, A., et al. (2008). Tasks, topics and relevance judging for the TREC Genomics Track: Five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval*. doi:10.1007/s10791-008-9072-x.

Seki, K., Costello, J., et al. (2004). TREC 2004 genomics track experiments at IUB. *The Thirteenth Text Retrieval Conference* (*TREC 2004*). Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/indianau-seki.geo.pdf.

Shatkay, H., Chen, N., et al. (2006). Integrating image data into biomedical text categorization. *Bioinformatics (Oxford, England), 22*, e446–e453. doi:10.1093/bioinformatics/btl235.

Stokes, N., Li, Y., et al. (2008). Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*. doi:10.1007/s10791-008-9073-9.

Strivens, M., & Eppig, J. (2004). Visualizing the laboratory mouse: Capturing phenotype information. *Genetica, 122*, 89–97. doi:10.1007/s10709-004-1435-7.

Venclovas, C., Zemla, A., et al. (2003). Assessment of progress over the CASP experiments. *Proteins, 53*, 585–595. doi:10.1002/prot.10530.

Voorhees, E., & Harman, D. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge, MA: MIT Press.

Wheeler, D., Barrett, T., et al. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 36*, D13–D21. doi:10.1093/nar/gkm1000.

Yeh, A., Hirschman, L., et al. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. *Bioinformatics (Oxford, England), 19*, I331–I339. doi:10.1093/bioinformatics/btg1046.

Zheng, Z., Brady, S., et al. (2005). Applying probabilistic thematic clustering for classification in the TREC 2005 genomics track. *The Fourteenth Text REtrieval Conference Proceedings* (*TREC 2005*). Gaithersburg, MD: National Institute for Standards and Technology. http://trec.nist.gov/pubs/trec14/papers/queensu.geo.pdf.