Report on TREC 2003 Genomics Track
First-Year Results and Future Plans

William Hersh
Track Chair
Oregon Health & Science University
Portland, OR, USA
hersh@ohsu.edu

The first year of the TREC Genomics Track (2003) was very successful, with a total of 29 groups participating. This made the track the second largest at TREC in terms of numbers of participating groups. In addition, we have been awarded a National Science Foundation Information Technology Research grant, which will provide five years of funding. Background on the motivation and evolution of the track can be found on the track Web site (http://medir.ohsu.edu/~genomics/) as well as in the Track Overview paper from the 2003 conference (http://medir.ohsu.edu/~genomics/overview.pdf).

In this report, we will discuss:
1. The 2003 track and its two tasks, with a focus on what approaches gave the best results.
2. The TREC 2003 meeting workshop and other planning for the 2004 track.
3. Next steps for the track.

First-Year Results

At the TREC 2003 meeting, the track had a plenary session which included an overview of track and presentations by four groups who were among the best-performing runs. However, we also selected presentations for the diversity of their approaches. Many other groups who participated had posters that described their techniques.

The primary task for the 2003 track was an ad hoc retrieval task using MEDLINE records (titles, abstracts, and human-assigned MeSH terms) as documents, gene names (and their synonyms) as queries, and Gene Reference into Function designations (GeneRIFs) as relevance judgments. GeneRIFs are annotations added to MEDLINE by indexers which describe the function of a gene (when an article describes it). They are not assigned exhaustively to all articles about the gene, but mainly to those deemed important by the indexers or others.

The main outcome measure for the primary task was mean average precision (MAP), though we calculated others, such as precision at X documents (where X was a range of different values) and R-prec (which measures precision at the retrieval set size of the number of relevant documents). A review of the papers, posters, and presentations allows some generalizations about which approaches led to good performance. Some analyses of the data yielded some additional insights.

The best-performing runs came from a research group (not affiliated with the operations of the library) from the National Library of Medicine. 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.

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

The teams from UC Berkeley and the National Research Council of Canada also achieved good results. Both of their approaches benefited from:

- Rules for recognizing gene name synonyms
- Filtering for organism name

The UC Berkeley approach included:

- A machine learning algorithm to classify documents likely to have GeneRIFs assigned to them
- Document ranking based on gene name occurrence rules

The NRC approach added:

- Unsupervised relevance feedback to find additional relevant articles
- 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 which included:

- 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 and Tarragon Consulting. No groups attempted modeled gene "function" in the sense of the GeneRIFs.

Approaches which 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. 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.

Some investigation of the test collection yielded some interesting knowledge. An analysis of relevance was done using output from the best OHSU training and test runs and confirmed what many suspected: GeneRIFs are imperfect as relevance judgments. While documents designated

as GeneRIFs are virtually certain to be relevant, many relevant documents are not GeneRIFs. Another group (Edinburg-Stanford) did some analysis exploring how MAP varied depending upon the 50 genes chosen for the queries. They found there was substantial variation in MAP depending on the random selection of genes for a given experiment. Most groups likewise noted substantial absolute differences between their training and test query results. This phenomenon is not necessarily surprising, and an interesting follow-up experiment would assess whether there were any relative changes in system performance across gene sets.

The secondary task also yielded interesting results, but of course was still clearly exploratory in nature. The goal of this task was to nominate the annotation text of the GeneRIF, with the gold standard being the text chosen by the NLM indexer. Preliminary analyses showed that the text often came sentences in the title or abstract of the MEDLINE record, with the title being used most commonly. In fact, just using the text of the titles alone achieved a baseline performance that few groups were able to outperform. The best approaches (Erasmus, Berkeley) used classifiers to rank sentences likely to contain the GeneRIF text.

Track workshop

An important part of the TREC meeting is the track workshop, where perspective from the current year can be discussed and planning for the following year and beyond can be done. The Genomics Track workshop had lively discussion and many good ideas for the future of the track.

A great deal of discussion focused on the development of tasks and queries that represented the information needs of real users. A number of individuals in the workshop came from biology backgrounds and wanted to make sure the track research would truly benefit the intended audience. It was also noted that while certain other groups of professionals information needs were well-studied (e.g., physicians, intelligence analysts), there was little systematic understanding of the needs of biology researchers. It was therefore decided that a smaller group should be convened to come up with a plan to investigate information needs in this domain, which could not only perform new research in this area in its own right, but also provide tasks and information needs for the track. This group has begun deliberating via email.

There was also discussion about the kinds of tasks we might want the track to have in the future. Certainly there will always be some sort of retrieval task; after all, this is TREC! But other ideas were put forth. There is still considerable interest in information extraction, although there is also another forum for extraction in biology, BioCreative (http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html). A number of other ideas were put forth:

- Information summarization, both in single and across multiple documents.
- Other information tasks, including curation/annotation and discovery of novelty and/or historical facts.
- Question-answering, similar to the TREC Question-Answering Track.

In addition, there was some deliberation over the information resources to be used in future years for the track. There remains a broad consensus that public domain resources, particularly the databases of the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) of the National Library of Medicine (NLM), should be the centerpiece. But there was also interest expressed in full-text journal articles (like those we had this year from Highwire Press) as well as the incorporation of other structured (non-text) data which is so plentiful in this domain. The was general agreement, however, that the selection of information resources should follow and not drive development of tasks and queries. So no decisions will be made on information resources until the information needs analysis is well underway.

Future Plans

A number of activities are planned to ensure that the 2004 track proceeds on schedule. The first activity will be a convening of those interested in characterizing information needs, which has already begun. There will also be a meeting at the Pacific Symposium on Biocomputing. In February, the track steering committee will meet and draw up a draft plan for the tasks, topics, documents, and relevance judging of the 2004 track. This plan will then be circulated to the track email list and finalized by the early spring. In the meantime, we will be preparing content (documents) for the test collection and addressing other logistical issues. Our goal will be to distribute the content and tasks/queries/topics by the end of the spring.