TREC 2004 Genomics Track Overview

William R. Hersh[1], Ravi Teja Bhuptiraju[1], Laura Ross[1], Phoebe Johnson[2], Aaron M. Cohen[1], Dale F. Kraemer[1]
[1]Oregon Health & Science University, Portland, OR, USA
[2]Biogen Idec Corp., Cambridge, MA

The TREC 2004 Genomics Track consisted of two tasks. The first task was a standard ad hoc retrieval task using topics obtained from real biomedical research scientists and documents from a large subset of the MEDLINE bibliographic database. The second task focused on categorization of full-text documents, simulating the task of curators of the Mouse Genome Informatics (MGI) system and consisting of three subtasks. One subtask focused on the triage of articles likely to have experimental evidence warranting the assignment of GO terms, while the other two subtasks focused on the assignment of the three top-level GO categories. The track had 33 participating groups.

1. Motivations and Background

The goal of the TREC Genomics Track is to create test collections for evaluation of information retrieval (IR) and related tasks in the genomics domain. The Genomics Track differs from all other TREC tracks in that it is focused on retrieval in a specific domain as opposed to general retrieval tasks, such as Web searching or question answering.

To date, the track has focused on advanced users accessing the scientific literature. The advanced users include biomedical scientists and database curators or annotators. New advances in biotechnologies have changed the face of biological research, particularly "high-throughput" techniques such as gene microarrays [1]. These not only generate massive amounts of data but also have led to an explosion of new scientific knowledge. As a result, this domain is ripe for improved information access and management.

The scientific literature plays a key role in the growth of biomedical research data and knowledge. Experiments identify new genes, diseases, and other biological processes that require further investigation. Furthermore, the literature itself becomes a source of "experiments" as researchers turn to it to search for knowledge that drives new hypotheses and research. Thus there are considerable challenges not only for better IR systems, but also for improvements in related techniques, such as information extraction and text mining [2].

Because of the growing size and complexity of the biomedical literature, there is increasing effort devoted to structuring knowledge in databases. The use of these databases is made pervasive by the growth of the Internet and Web as well as a commitment of the research community to put as much data as possible into the public domain. Figure 1 depicts the overall process of "funneling" the literature to structure knowledge, showing the information system tasks used at different levels along the way. This figure shows our view of the optimal uses for IR and the related areas of information extraction and text mining.

One of the many key efforts is to annotate the function of genes. To facilitate this, the research community has come together to develop the Gene Ontology (GO, www.geneontology.org) [3]. While the GO is not an ontology in the purists' sense, it is a large, controlled vocabulary based on three axes or hierarchies:

- Molecular function - the activity of the gene product at the molecular (biochemical) level, e.g. protein binding
- Biological process - the biological activity carried out by the gene process, e.g., cell differentiation
- Cellular component - where in the cell the gene product functions, e.g., the nucleus

A major use of the GO has been to annotate the genomes of organisms used in biological research. The annotations are often linked to other information, such as literature, the gene sequence, the structure of the resulting protein, etc.. An increasingly common approach is to develop "model organism databases" that bring together all this information in an easy to use format. Some of the better known model organism databases include those devoted to the mouse (Mouse Genome Informatics, MGI,

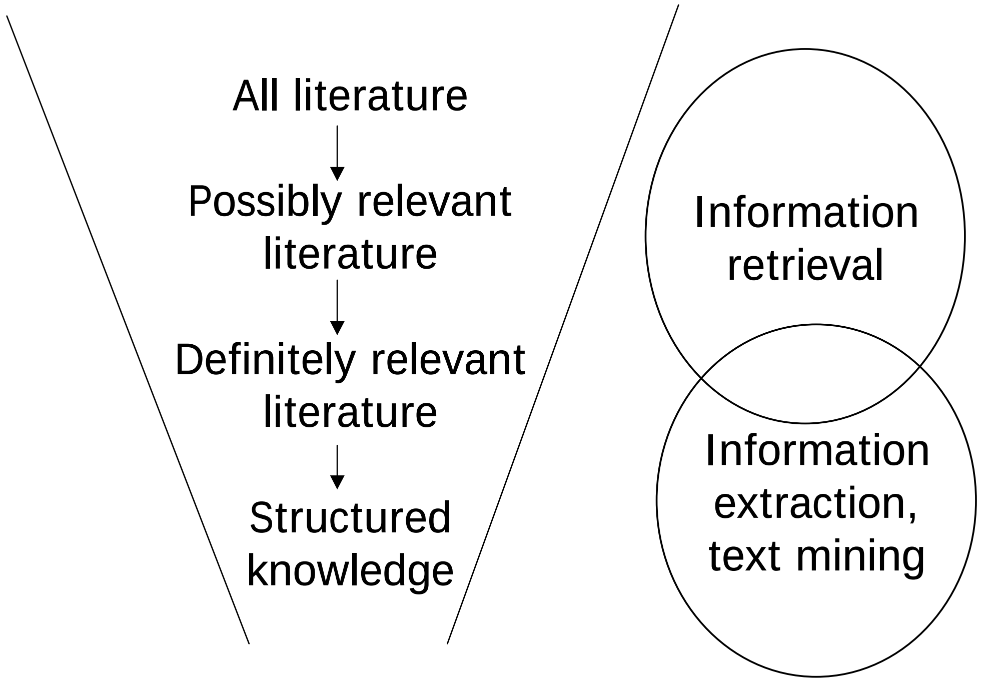
All literature
Possibly relevant literature
Definitely relevant literature
Structured knowledge
Information retrieval
Information extraction, text mining

Figure 1 - The steps in deriving knowledge from the biomedical literature and the associated information systems used along the way.

www.informatics.jax.org) and the yeast (Saccharomyces Genome Database, SGD, www.yeastgenome.org). These databases require extensive human effort for annotation or curation, which is usually done by PhD-level researchers.

These curators could be aided substantially by high-quality information tools, including IR systems.

The 2004 track was the second year of the TREC Genomics Track. This year was different from the first year, as we had resources available to us from a National Science Foundation (NSF) Information Technology Research (ITR) grant that allowed for programming support and relevance judgments. In contrast, for the 2003 track we had to rely on proxies for relevance judgments and other gold standard data [4].

The Genomics Track is overseen by a steering committee of individuals with a background in IR and/or genomics. In early 2003, the committee produced a “road map” that called for modifying one experimental “facet” each year. For the purposes of the roadmap (based on the NSF grant proposal), the original year (2003) was Year 0, making 2004 Year 1. The original plan was to add new types of content in Year 1 and new types of information needs in Year 2. Because we were unable to secure substantial numbers of full text documents for the ad hoc retrieval task in 2004, we decided to reverse the order of the roadmap for Years 1 and 2. This meant we focused on new types of information needs for 2004 (and hopefully new types of content in 2005). However, it should be noted that even in this era of virtually all biomedical journals being available electronically, most users of the literature start their searches using MEDLINE.

2. Overview of Track

In TREC 2004, the Genomics Track had two tasks, the second of which was subdivided into subtasks. The first task was a standard ad hoc retrieval task using topics obtained from surveying real research scientists and searching in a large subset of the MEDLINE bibliographic database. The second task focused on categorization of full-text documents, simulating the task of curators for the MGI system. One subtask focused on the triage of articles likely to have experimental evidence warranting the assignment of GO terms, while the other two subtasks focused on the assignment of the three GO categories (indicating the assignment of a term within them).

A total of 145 runs were submitted for scoring. There were 47 runs from 27 groups submitted for the ad hoc task. There were 98 runs submitted from 20 groups for the categorization task. These were distributed across the subtasks of the categorization task as follows: 59 for the triage subtask, 36 for the annotation hierarchy subtask, and three for the annotation hierarchy plus evidence code subtask. A total of 33 groups participated in the 2004 Genomics Track, making it the track with the most participants in all of TREC 2004.

The data are currently available to track participants on password-protected Web sites but will be made available to non-TREC participants in early 2005. The version of data released in early 2005 will be updated to correct some minor errors associated with the official TREC 2004 data.

3. Ad Hoc Retrieval Task

The goal of the ad hoc task was to mimic conventional searching. The use case was a scientist with a specific information need, searching the MEDLINE bibliographic database to find relevant articles to retrieve.

3.1 Documents

The document collection for the ad hoc retrieval task was a 10-year subset of MEDLINE. We contemplated the use of full-text documents in this task but were unable to procure an adequate amount to represent real-world searching. As such, we chose to use MEDLINE. As noted above, however, despite the widespread availability of on-line, full-text scientific journals at present, most searchers of the biomedical literature still use MEDLINE as an entry point. Consequently, there is great value in being able to search MEDLINE effectively.

The subset of MEDLINE used for the track 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. We used the DCOM field and not the Date Published (DP). As a result, some records were published but not completed prior to 1994, i.e., the collection had:

- 2,814 (0.06%) DPs prior to 1980
- 8,388 (0.18%) DPs prior to 1990
- 138,384 (3.01%) DPs prior to 1994

The remaining 4,452,624 (96.99%) DPs were within the 10 year period of 1994-2004.

The data was made available in two formats:

- MEDLINE - the standard NLM format in ASCII text with fields indicated and delimited by 2-4 character abbreviations (uncompressed - 9,587,370,116 bytes, gzipped - 2,797,589,659 bytes)
- XML - the newer NLM XML format (uncompressed - 20,567,278,551 bytes, gzipped - 3,030,576,659 bytes)

3.2 Topics

The topics for the ad hoc retrieval task were developed from the information needs of real biologists and modified as little as possible to create needs statements with a reasonable estimated amount of relevant articles (i.e., more than zero but less than one thousand). The information needs capture began with interviews by 12 volunteers who sought biologists in their local environments. A total of 43 interviews yielded 74 information needs. Some of these volunteers, as well as an additional four individuals, created topics in the proposed format from the original interview data. We aimed to have each information need reviewed more than once but were only able to do this with some, ending up with a total of 91 draft topics. The same individuals then were assigned different draft topics for searching on PubMed so they could be modified to generate final topics with a reasonable number of relevant articles. The track chair made one last pass to make the formatting consistent and extract the 50 that seemed most suitable as topics for the track.

The topics were formatted in XML and had the following fields:

- ID - 1 to 50
- Title - abbreviated statement of information need
- Information need - full statement information need
- Context - background information to place information need in context

We created an additional five “sample” topics, one of which is displayed in Figure 2.

```
<TOPIC>
  <ID>51</ID>
  <TITLE>pBR322 used as a gene vector</TITLE>
  <NEED>Find information about base sequences and restriction maps in plasmids that are used
          as gene vectors.</NEED>
  <CONTEXT>The researcher would like to manipulate the plasmid by removing a particular
          gene and needs the original base sequence or restriction map information of the
          plasmid.</CONTEXT>
  </TOPIC>
```

Figure 2 - Sample topic for ad hoc retrieval task.

3.3 Relevance Judgments

Relevance judgments were done using the conventional “pooling method” whereby a fixed number of top-ranking documents from each official run were pooled and provided to an individual (blinded to the number of groups who retrieved the document and what their search statements were). The relevance assessor then judged each document for the specific topic query as definitely relevant (DR), possibly relevant (PR), or not relevant (NR). A subset of documents were also judged in duplicate to assess interjudge reliability using the kappa measure [5]. For the official results, which required binary relevance judgments, documents that were rated DR or PR were considered relevant.

The pools were built as follows. Each of the 27 groups designated a top-precedence run that would be used for relevance judgments, typically what they thought would be their best-performing run. We took, on average, the top 75 documents for each topic from these 27 runs and eliminated the duplicates to create a single pool for each topic. The average pool size (average number of documents judged per topic) was 976, with a range of 476-1450.

The judgments were done by two individuals with backgrounds in biology. One was a PhD biologist and the other an undergraduate biology student. Table 1 shows the pool size and number of relevant documents for each topic. (It also shows the overall results, to be described later.)

For the kappa measurements, we selected every tenth article from six topics. As each judge had already judged the documents for three of the topics, we compared these extra judgments with the regular ones done by the other judge. The results of the duplicate judgments are shown in Table 2. The resulting kappa score was 0.51, indicating a “fair” level of agreement but not being too different from similar relevance judgment activities in other domains, e.g., [6]. In general, the PhD biologist assigned more articles in the relevant category than the undergraduate.

3.4 Evaluation Measures

The primary evaluation measure for the task was mean average precision (MAP). Results were calculated using the trec_eval program, a standard scoring system for TREC. A statistical analysis was performed using a repeated measures analysis of variance, with posthoc Tukey tests for pairwise comparisons. In addition to analyzing MAP, we also assessed precision at 10 and 100 documents.

3.5 Results

The results of all participating groups are shown in Table 3. The statistical analysis for MAP demonstrated significance across all the runs, with the pairwise significance for the top run (pllsgen4a2) not obtained until the run RMITa about one-quarter of the way down the results.

The best official run was achieved by Patolis Corp. [7]. This run used 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, blind relevance feedback (also known as blind query expansion), and use of all three fields in the query. This group also reported a post-submission run that added the language modeling technique of Dirichlet-Prior smoothing to achieve an even higher MAP of 0.4264.

Table 1 - Ad hoc retrieval topics, number of relevant documents, and average results for all runs.

Topic	**Pool**	**Definitely Relevant**	**Possibly Relevant**	**Not Relevant**	**D & P Relevant**	**MAP average**	**P@10 average**	**P@100 average**
1	879	38	41	800	79	0.3073	0.7383	0.2891
2	1264	40	61	1163	101	0.0579	0.2787	0.1166
3	1189	149	32	1008	181	0.0950	0.3298	0.2040
4	1170	12	18	1140	30	0.0298	0.0894	0.0360
5	1171	5	19	1147	24	0.0564	0.1340	0.0349
6	787	41	53	693	94	0.3993	0.8468	0.3938
7	730	56	59	615	115	0.2006	0.4936	0.2704
8	938	76	85	777	161	0.0975	0.3872	0.2094
9	593	103	12	478	115	0.6114	0.7957	0.6196
10	1126	3	1	1122	4	0.5811	0.2532	0.0277
11	742	87	24	631	111	0.3269	0.5894	0.3843
12	810	166	90	554	256	0.4225	0.7234	0.5866
13	1118	5	19	1094	24	0.0288	0.1021	0.0274
14	948	13	8	927	21	0.0479	0.0894	0.0270
15	1111	50	40	1021	90	0.1388	0.2915	0.1800
16	1078	94	53	931	147	0.1926	0.4489	0.2883
17	1150	2	1	1147	3	0.0885	0.0511	0.0115
18	1392	0	1	1391	1	0.6254	0.0660	0.0072
19	1135	0	1	1134	1	0.1594	0.0362	0.0062
20	814	55	61	698	116	0.1466	0.3957	0.2238
21	676	26	54	596	80	0.2671	0.4702	0.2796
22	1085	125	85	875	210	0.1354	0.4234	0.2709
23	915	137	21	757	158	0.1835	0.3745	0.2747
24	952	7	19	926	26	0.5970	0.7468	0.1685
25	1142	6	26	1110	32	0.0331	0.1000	0.0330
26	792	35	12	745	47	0.4401	0.7298	0.2411
27	755	19	10	726	29	0.2640	0.4319	0.1355
28	836	6	7	823	13	0.2031	0.2532	0.0643
29	756	33	10	713	43	0.1352	0.1809	0.1515
30	1082	101	64	917	165	0.2116	0.4872	0.3113
31	877	0	138	739	138	0.0956	0.2489	0.2072
32	1107	441	55	611	496	0.1804	0.6085	0.4787
33	812	30	34	748	64	0.1396	0.2234	0.1647
34	778	1	30	747	31	0.0644	0.0830	0.0668
35	717	253	18	446	271	0.3481	0.8213	0.6528
36	676	164	90	422	254	0.4887	0.7638	0.6700
37	476	138	11	327	149	0.5345	0.7426	0.6564
38	1165	334	89	742	423	0.1400	0.5915	0.4043
39	1350	146	171	1033	317	0.0984	0.3936	0.2689
40	1168	134	143	891	277	0.1080	0.3936	0.2796
41	880	333	249	298	582	0.3356	0.6766	0.6521
42	1005	191	506	308	697	0.1587	0.6596	0.5702
43	739	25	170	544	195	0.1185	0.6915	0.2553
44	1224	485	164	575	649	0.1323	0.6149	0.4632
45	1139	108	48	983	156	0.0286	0.1574	0.0711
46	742	111	86	545	197	0.2630	0.7362	0.4981
47	1450	81	284	1085	365	0.0673	0.3149	0.2355
48	1121	53	102	966	155	0.1712	0.4021	0.2557
49	1100	32	41	1027	73	0.2279	0.5404	0.2049
50	1091	79	223	789	302	0.0731	0.3447	0.2534
Mean	975.1	92.6	72.8	809.7	165.4	0.2171	0.4269	0.2637
Median	978.5	54	44.5	783	115.5	0.1590	0.3989	0.2472
Min	476	0	1	298	1	0.0286	0.0362	0.0062
Max	1450	485	506	1391	697	0.6254	0.8468	0.6700

Table 2 - Kappa results for interjudge agreement in relevant judgments for ad hoc retrieval task.

Judge 2	Definitely relevant	Possibly relevant	Not relevant	Total
Judge 1				
Definitely relevant	62	35	8	105
Possibly relevant	11	11	5	27
Not relevant	14	57	456	527
Total	87	103	469	659

The next best run was achieved by the University of Waterloo [8]. This group used a variety of approaches including Okapi weighting, blind relevance feedback, and various forms of domain-specific query expansion. Their blind relevance feedback made use of usual document feedback as well as feedback from passages. Their domain-specific query expansion included expanding lexical variants as well as expanding acronym, gene, and protein name synonyms.

A number of groups used boosting of word weights in queries or documents. Tsinghua University boosted words in titles and abstracts, along with using blind query expansion [9]. Alias-i Corp. boosted query words in the title and need statements [10]. University of Tampere found value in identifying and using bi-gram phrases [11].

A number of groups implemented techniques, however, that were detrimental. This is evidenced by the OHSU runs, which used the Lucene system "out of the box" that applies TF*IDF weighting [12]. Approaches that attempted to map to controlled vocabulary terms did not fare as well, such as Indiana University [13], University of California Berkeley [14], and the National Library of Medicine [15]. 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 which techniques provided benefit. Figure 3 shows the official results graphically with annotations for the first run statistically significant from the top run as well as the OHSU "baseline."

As typically occurs in TREC ad hoc runs, there was a great deal of variation within individual topics, as is seen in Table 1. Figure 4 shows the average MAP across groups for each topic. Figure 5 presents the same data sorted to give a better indication of the variation across topics. There was a fairly strong relationship between the average and maximum MAP for each topic (Figure 6), while the number of relevant per topic versus MAP was less associated (Figure 7).

4. Categorization Task

In the categorization task, we simulated two of the classification activities carried out by human annotators for the MGI system: a triage task and two simplified variations of MGI's annotation task. Systems were required to classify full-text documents from a two-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 GO codes about gene function with another code describing the type of experimental evidence supporting assignment of the GO code. The huge amount of literature requiring curation creates a challenge for MGI, as their 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.

Table 3 - Ad hoc retrieval results, sorted by mean average precision.

Run	**Group (reference)**	**Manual/ Automatic**	**Mean Average Precision**	**Relevant at 10 documents**	**Relevant at 100 documents**
pllsgen4a2	patolis.fujita [7]	A	0.4075	6.04	41.96
uwmtDg04tn	u.waterloo.clarke [8]	A	0.3867	6.24	42.1
pllsgen4a1	patolis.fujita [7]	A	0.3689	5.7	39.36
THUIRgen01	tsinghua.ma [9]	M	0.3435	5.82	39.24
THUIRgen02	tsinghua.ma [9]	A	0.3434	5.94	39.44
utaauto	u.tampere [11]	A	0.3324	5.02	32.26
uwmtDg04n	u.waterloo.clarke [8]	A	0.3318	5.68	36.84
PSE	german.u.cairo [18]	A	0.3308	5.86	36.66
tnog3	tno.kraaij [19]	A	0.3247	5.6	36.56
tnog2	tno.kraaij [19]	A	0.3196	5.62	36.04
utamanu	u.tampere [11]	M	0.3128	6.52	38.88
aliasiBase	alias-i [10]	A	0.3094	5.38	34.58
ConversManu	converspeech [20]	M	0.2931	5.82	37.18
RMITa	rmit.scholer [21]	A	0.2796	5.12	31.4
aliasiTerms	alias-i [10]	A	0.2656	4.8	30.3
akoike	u.tokyo (none)	M	0.2427	4.48	31.3
OHSUNeeds	ohsu.hersh [12]	A	0.2343	3.84	26.46
tgnSplit	tarragon [22]	A	0.2319	4.86	29.26
UIowaGN1	u.iowa [23]	A	0.2316	4.76	28.5
tq0	nlm.umd.ul [15]	A	0.2277	5.12	30.1
OHSUAll	ohsu.hersh [12]	A	0.2272	4.32	27.76
LHCUMDSE	nlm.umd.ul [15]	A	0.2191	3.9	24.18
akoyama	u.tokyo (none)	M	0.2155	4.52	25.62
PDTNsmp4	u.padova [24]	A	0.2074	4.56	23.18
PD50501	u.padova [24]	A	0.2059	4.42	25.18
RMITb	rmit.scholer [21]	A	0.2059	4.56	27.26
UBgtNormJM1	suny.buffalo [25]	A	0.2043	4.34	25.38
ConversAuto	converspeech [20]	A	0.2013	3.88	22.8
york04g2	york.u [26]	M	0.2011	5.5	25.8
tgnNecaux	tarragon [22]	A	0.1951	4.08	23.58
lga1	indiana.u.seki [13]	A	0.1833	3.08	22.86
york04g1	york.u [26]	A	0.1794	4.14	26.96
lga2	indiana.u.seki [13]	A	0.1754	3.1	20.22
rutgersGAH1	rutgers.dayanik [16]	A	0.1702	4.66	26.76
wdvqlxa1	indiana.u.yang [27]	A	0.1582	4.2	24.78
wdvqlx1	indiana.u.yang [27]	A	0.1569	4.26	24.26
DCUmatn1	dubblincity.u [28]	M	0.1388	3.28	17.84
BioTextAdHoc	u.cberkeley.hearst [14]	A	0.1384	3.76	23.76
shefauto2	u.sheffield.gaizauskas [29]	A	0.1304	3.66	18.5
rutgersGAH2	rutgers.dayanik [16]	A	0.1303	3.42	19.48
shefauto1	u.sheffield.gaizauskas [29]	A	0.1294	3.54	18.92
run1	utwente (none)	M	0.1176	1.5	10.5
MeijiHilG	meiji.u [30]	A	0.0924	2.1	15.24
DCUma	dubblincity.u [28]	M	0.0895	2.4	15.46
csusm	u.sanmarcos [31]	M	0.0123	0.44	1.6
edinauto2	u.edinburgh.sinclair [32]	A	0.0017	0.46	1.6
edinauto5	u.edinburgh.sinclair [32]	A	0.0012	0.36	1.3
Mean			0.2074	4.48	26.46

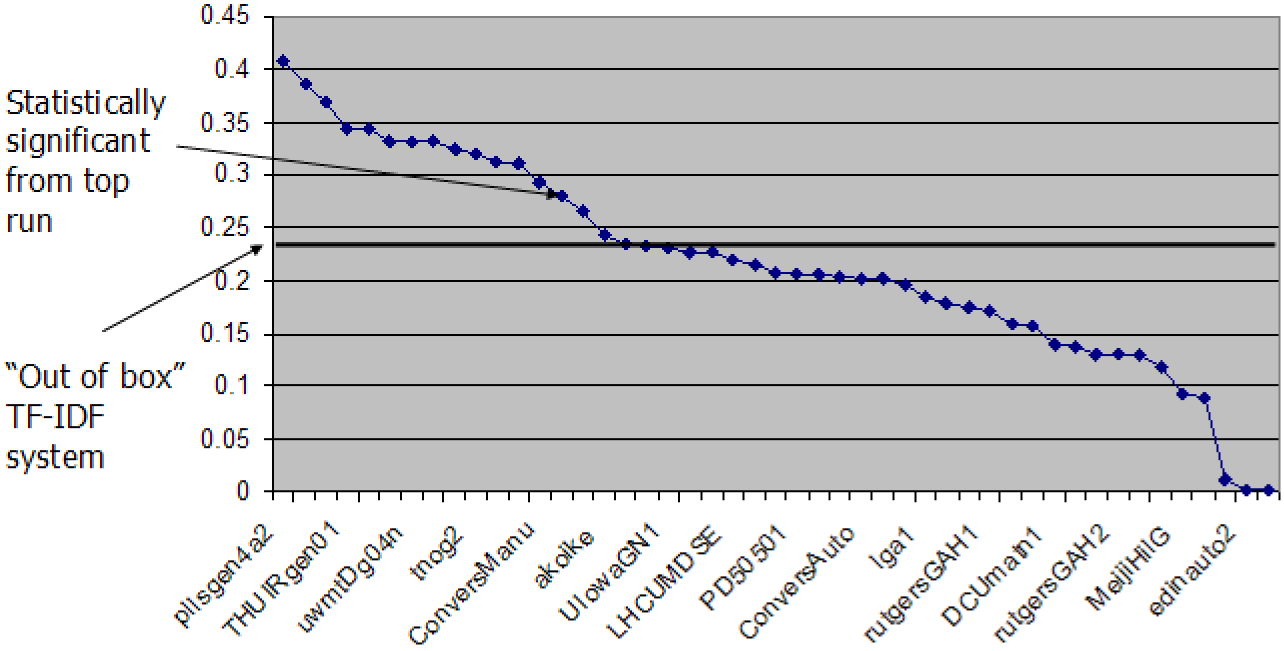

Figure 3 - Ad hoc retrieval runs sorted by MAP score. The highest run to obtain statistical significance (RMITa) from the top run (pllsgen4a2) is denoted, along with the “out of the box” TF*IDF run (OHSUNeeds) are annotated.

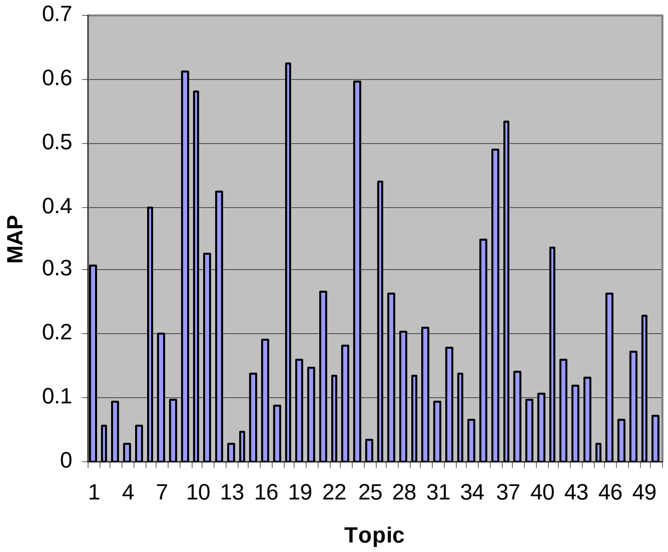

Figure 4 - MAP by topic for the ad hoc task.

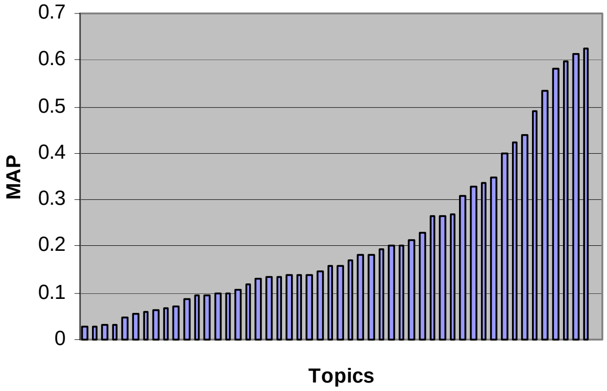

Figure 5 - MAP by topic for the ad hoc task sorted by MAP.

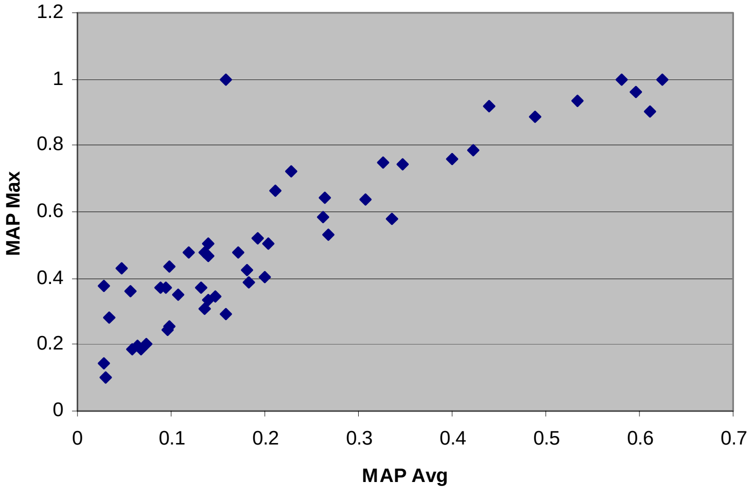

Figure 6 - The maximum MAP plotted vs. average MAP for the ad hoc retrieval task runs.

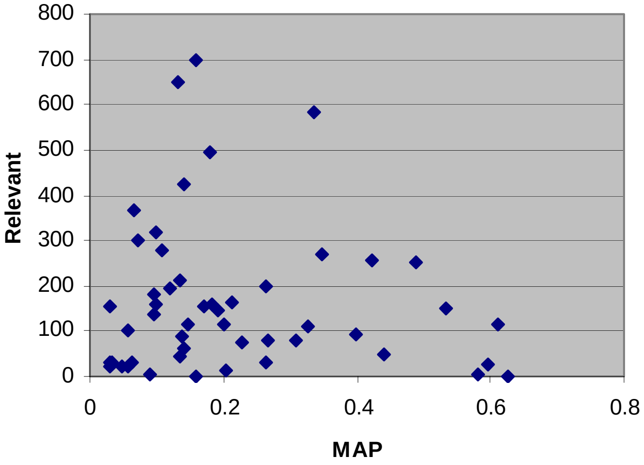

Figure 7 - The number of relevant per topic plotted vs. MAP for the ad hoc retrieval task.

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. Our triage task involved correctly classifying which documents had been selected for GO annotation in this process.
3. Annotation - The third step is the actual curation with GO terms. Curators identify genes for which there is experimental evidence to warrant assignment of GO codes. Those GO codes are assigned, along with a code for each indicating the type of experimental evidence. There can more than one gene assigned GO codes in a given paper and there can be more than one GO code assigned to a gene. In general, and in our collection, there is only one evidence code per GO code assignment per paper. Our annotation task involved a modification of this annotation step as described below.

4.1 Documents

The documents for the categorization task consisted of articles from three journals over two years, reflecting the full-text documents we were able to obtain from Highwire Press (www.highwire.org). Highwire is a "value added" electronic publisher of scientific journals. Most journals in their collection are published by professional associations, with the copyright remaining with the associations. Highwire originally began with biomedical journals, but in recent years has expanded into other disciplines. They have also supported IR and related research by acting as an intermediary between consenting publishers and information systems research groups who want to use their journals, such as the Genomics Track.

The journals available and used by our track this year were *Journal of Biological Chemistry* (JBC), *Journal of Cell Biology* (JCB), and *Proceedings of the National Academy of Science* (PNAS). These journals have a good proportion of mouse genome articles. Each of the papers from these journals was provided in SGML format based on Highwire's Document Type Definition (DTD). We used articles from the year 2002 for training data and from 2003 for test data. The documents for the categorization tasks came from a subset of articles having the words *mouse*, *mice* or *murine* as described above. We created a crosswalk file (look-up table) that matched an identifier for each Highwire article (its file name) and its corresponding PubMed ID (PMID). Table 4 shows the total number of articles in each journal and the number in each journal included in subset used by the track. The SGML training document collection was 150 megabytes in size compressed and 449 megabytes uncompressed. The SGML test document collection was 140 megabytes compressed and 397 megabytes uncompressed.

Since MGI annotation lags behind article publication, a not insubstantial number of papers have been selected for annotation but not yet annotated. From the standpoint of the triage subtask, we wanted to use all of these articles as positive examples, since they all were selected for GO annotation. However, we could not use the articles not yet annotated for the annotation hierarchy task, since we did not have the annotations. We also needed a set of negative examples for the annotation hierarchy task and chose to use articles selected for action by MGI for other (i.e., non-GO annotation) actions. Figure 8 shows the groups of documents and how they were assigned into being positive and negative examples for the subtasks.

4.2 Triage Subtask

The goal of this task was to correctly identify papers that were deemed to have experimental evidence warranting annotation with GO codes. Positive examples included papers designated for GO annotation by MGI. As noted above, some of these papers had not yet been annotated. Negative examples were all papers not designated for GO annotation in the operational MGI system. For the training data (2002), there were 375 positive examples, meaning that there were 5837-375 = 5462 negative examples. For the test data (2003), there were 420 positive examples, meaning that there were 6043-420 = 5623 negative examples. It should also be noted that the MGI system is, like most operational databases, continuously updated, so the data for the track represented a snapshot of the database obtained in May, 2004. (As described later, an updated version of the data will be available in 2005.)

Table 4 - Number of papers total and available in the *mouse, mus,* or *murine* subset.

Journal	**2002 papers - total, subset**	**2003 papers - total, subset**	**Total papers - total, subset**
JBC	6566, 4199	6593, 4282	13159, 8481
JCB	530, 256	715, 359	1245, 615
PNAS	3041, 1382	2888, 1402	5929, 2784
Total papers	10137, 5837	10196, 6043	20333, 11880

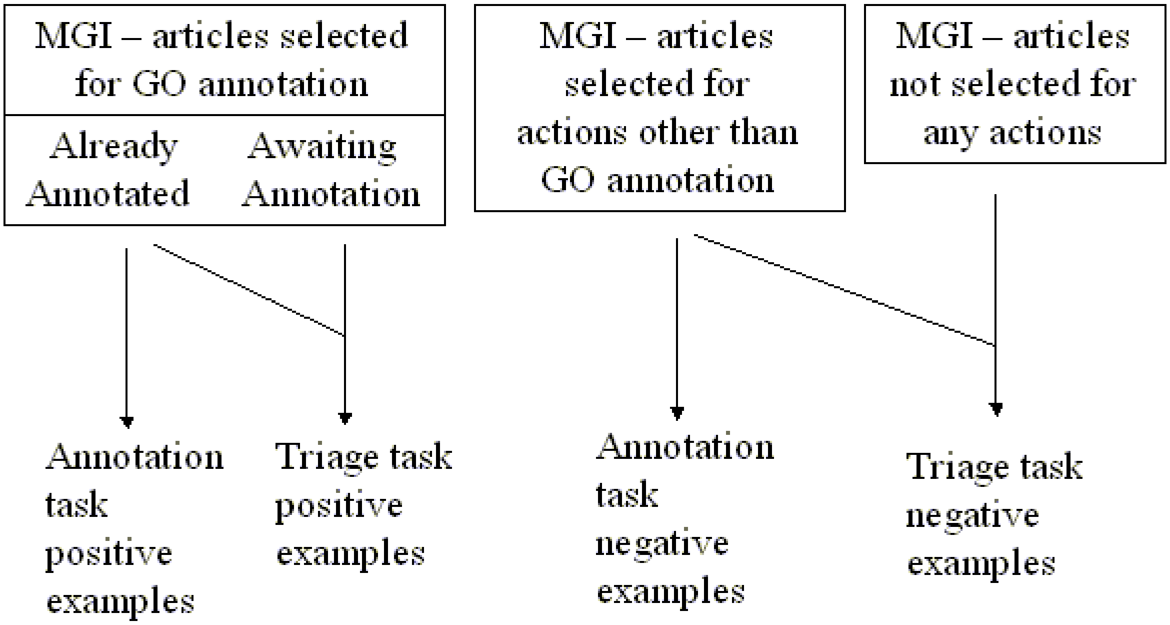
MGI – articles selected for GO annotation
Already Annotated
Awaiting Annotation
MGI – articles selected for actions other than GO annotation
MGI – articles not selected for any actions
Annotation task positive examples
Triage task positive examples
Annotation task negative examples
Triage task negative examples

Figure 8 - Grouping of documents for categorization subtasks.

The evaluation measure for the triage task was the utility measure often applied in text categorization research and used by the former TREC Filtering Track. This measure contains coefficients for the utility of retrieving a relevant and retrieving a nonrelevant document. We used a version that was normalized by the best possible score:

$$U_{norm} = U_{raw} / U_{max}$$

where U_{norm} was the normalized score, U_{raw} the raw score, and U_{max} the best possible score.

The coefficients for the utility measure were derived as follows. For a test collection of documents to categorize, U_{raw} is calculated as:
$U_{raw} = (u_r * \text{relevant-docs-retrieved}) + (u_{nr} * \text{nonrelevant-docs-retrieved})$
where:

- u_r = relative utility of relevant document
- u_{nr} = relative utility of nonrelevant document

We used values for u_r and u_{nr} that were driven by boundary cases for different results. In particular, we wanted (thought it was important) the measure to have the following characteristics:

- Completely perfect prediction - $U_{norm} = 1$
- All documents designated positive (triage everything) - $1 > U_{norm} > 0$
- All documents designated negative (triage nothing) - $U_{norm} = 0$
- Completely imperfect prediction - $U_{norm} < 0$

In order to achieve the above boundary cases, we had to set $u_r > 1$. The ideal approach would have been to interview MGI curators and use decision-theoretic approaches to determine their utility. However, time constraints did not allow this. Deciding that the triage-everything approach should have a higher score than the triage-nothing approach, we estimated that a U_{norm} in the range of 0.25-0.3 for the triage-everything condition would be appropriate. Solving for the above boundary cases with U_{norm} ~ 0.25-0.3 for that case, we obtained a value for u_r ~ 20. To keep calculations simple, we choose a value of u_r = 20. Table 5 shows the value of U_{norm} for the boundary cases.

The measure U_{max} was calculated by assuming all relevant documents were retrieved and no nonrelevant documents were retrieved, i.e., U_{max} = u_r * all-relevant-docs-retrieved.

Thus, for the training data,

U_{raw} = (20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved

U_{max} = 20 * 375 = 7500

U_{norm} = [(20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved] / 7500

Likewise, for the test data,

U_{raw} = (20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved

U_{max} = 20 * 420 = 8400

U_{norm} = [(20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved] / 8400

The results of the triage subtask are shown in Table 6. A variety of groups used classifiers based on machine learning techniques. The higher scoring runs tended to make use of MeSH terms in some fashion. The best performing run came from Rutgers University, using the MEDLINE record, weighting, and filtering by the MeSH term *Mice* [16]. They achieved a U_{norm} of 0.6512. However, this group also noted that the MeSH term *Mice* alone scored better than all but the single top run, with a U_{norm} of 0.6404. This meant that no other approach was better able to classify documents for triage than simply using the MeSH term *Mice* from the MEDLINE record. Of course, this run only achieved a recall of about 15% (with a recall of 89%), so this feature is far from a perfect predictor. In an another analysis of the data, Cohen noted that there was conceptual drift across the collection, with the features identified as strong predictors in the training data not necessarily continuing to be strong predictors in the test data [12]. All of the triage subtask results are shown graphically in Figure 9, along with the utility for the MeSH term *Mice* and the decision to select all articles.

4.3 Annotation Subtask

The primary goal of this task was, given an article and gene name, to correctly identify which of the GO hierarchies (also called domains) had terms within them that were annotated by the MGI curators. Note that the goal of this task was not to select the actual GO term, but rather to select the one or more GO hierarchies (molecular function, biological process, or cellular component) from which terms had been selected to annotate the gene for the article. Papers that were annotated had terms from one to three hierarchies.

For negative examples, we used 555 papers that had a gene name assigned but were used for other purposes by MGI. As such, these papers had no GO annotations. These papers did, however, have one or more gene assigned by MGI for the other annotation purposes.

A secondary subtask was to identify the correct GO evidence code that went with the hierarchy code. Only two groups took part in this subtask.

Table 7 shows the contents and counts of the data files for this subtask. For the training data, there were a total of 504 documents that were either positive (one or more GO terms assigned) or negative (no GO terms assigned) examples. From these documents, a total of 1291 genes had been assigned by MGI. (The Genes file contained the MGI identifier, the gene symbol, and the gene name. It did not contain any other synonyms.) There were 1418 unique possible document-gene pairs in the training data. The data from the first three rows of Table 7 differ from the rest in that they contained data merged from positive and negative examples. These were what would be used as input for systems to nominate GO domains or the GO domains plus their evidence codes per the annotation task. When the test data were released, these three files were the only ones that were provided.

For the positive examples in the training data, there were 178 documents and 346 document-gene pairs. There were 589 document-gene name-GO domain tuples (out of a possible 346 * 3 = 1038). There were 640 document-gene name-GO domain-evidence code tuples. A total of 872 GO plus evidence codes had been assigned to these documents. For the negative examples, there were 326 documents and 1072 document-gene pairs. This meant that systems could possibly assign 1072*3 = 3216 document-gene name-GO domain tuples.

Table 5 - Boundary cases for utility measure of triage task for training and test data.

Situation	**U_{norm} - Training**	**U_{norm} - Test**
Completely perfect prediction	1.0	1.0
Triage everything	0.27	0.33
Triage nothing	0	0
Completely imperfect prediction	-0.73	-0.67

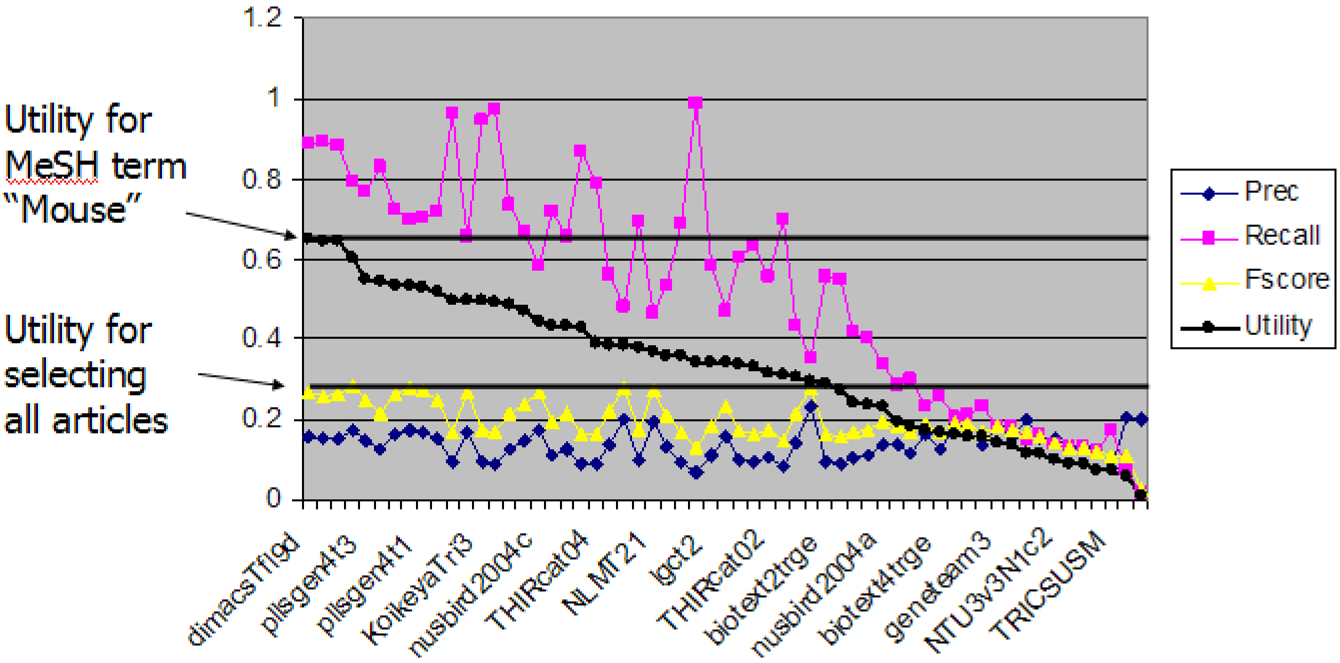

Figure 9 - Triage subtask runs sorted by U_{norm} score. The U_{norm} for the MeSH term *Mice* as well as for selecting all articles as positive is shown.

The evaluation measures for the annotation subtasks were based on the notion of identifying tuples of data. Given the article and gene, systems designated one or both of the following tuples:

- <article, gene, GO hierarchy code>
- <article, gene, GO hierarchy code, evidence code>

We employed a global recall, precision, and F measure evaluation measure for each subtask:

Recall = number of tuples correctly identified / number of correct tuples

Precision = number of tuples correctly identified / number of tuples identified

F = (2 * recall * precision) / (recall + precision)

For the training data, the number of correct <article, gene, GO hierarchy code> tuples was 589, while the number of correct <article, gene, GO hierarchy code, evidence code> tuples was 640.

The annotation hierarchy subtask results are shown in Table 8, while the annotation hierarchy subtask plus evidence code results are shown in Table 9. As noted above, the primary evaluation measure for this task was the F-score. Due to their only being a single measure per run, we were unable to perform comparative statistics. Figure 10 shows the annotation hierarchy subtask results graphically.

Table 6 - Triage subtask runs, sorted by utility.

Run	**Group (reference)**	**Precision**	**Recall**	**F-score**	**Utility**
dimacsTfl9d	rutgers.dayanik [16]	0.1579	0.8881	0.2681	0.6512
dimacsTl9mhg	rutgers.dayanik [16]	0.1514	0.8952	0.259	0.6443
dimacsTfl9w	rutgers.dayanik [16]	0.1553	0.8833	0.2642	0.6431
dimacsTl9md	rutgers.dayanik [16]	0.173	0.7952	0.2841	0.6051
pllsgen4t3	patolis.fujita [7]	0.149	0.769	0.2496	0.5494
pllsgen4t4	patolis.fujita [7]	0.1259	0.831	0.2186	0.5424
pllsgen4t2	patolis.fujita [7]	0.1618	0.7238	0.2645	0.5363
pllsgen4t5	patolis.fujita [7]	0.174	0.6976	0.2785	0.532
pllsgen4t1	patolis.fujita [7]	0.1694	0.7024	0.273	0.5302
GUCwdply2000	german.u.cairo [18]	0.151	0.719	0.2496	0.5169
KoikeyaTri1	u.tokyo (none)	0.0938	0.9643	0.171	0.4986
OHSUVP	ohsu.hersh [12]	0.1714	0.6571	0.2719	0.4983
KoikeyaTri3	u.tokyo (none)	0.0955	0.9452	0.1734	0.4974
KoikeyaTri2	u.tokyo (none)	0.0913	0.9738	0.167	0.4893
NLMT2SVM	nlm.umd.ul [15]	0.1286	0.7333	0.2188	0.4849
dimacsTl9w	rutgers.dayanik [16]	0.1456	0.6643	0.2389	0.4694
nusbird2004c	mlg.nus [33]	0.1731	0.5833	0.267	0.444
lgct1	indiana.u.seki [13]	0.1118	0.7214	0.1935	0.4348
OHSUNBAYES	ohsu.hersh [12]	0.129	0.6548	0.2155	0.4337
NLMT2BAYES	nlm.umd.ul [15]	0.0902	0.869	0.1635	0.4308
THIRcat04	tsinghua.ma [9]	0.0908	0.7881	0.1628	0.3935
GUClin1700	german.u.cairo [18]	0.1382	0.5595	0.2217	0.3851
NLMT22	nlm.umd.ul [15]	0.1986	0.481	0.2811	0.3839
NTU2v3N1	ntu.chen [34]	0.1003	0.6905	0.1752	0.381
NLMT21	nlm.umd.ul [15]	0.195	0.4643	0.2746	0.3685
GUCply1700	german.u.cairo [18]	0.1324	0.5357	0.2123	0.3601
NTU3v3N1	ntu.chen [34]	0.0953	0.6857	0.1673	0.3601
NLMT2ADA	nlm.umd.ul [15]	0.0713	0.9881	0.133	0.3448
lgct2	indiana.u.seki [13]	0.1086	0.581	0.183	0.3426
GUClin1260	german.u.cairo [18]	0.1563	0.469	0.2345	0.3425
THIRcat01	tsinghua.ma [9]	0.1021	0.6024	0.1746	0.3375
NTU4v3N1416	ntu.chen [34]	0.0948	0.6357	0.165	0.3323
THIRcat02	tsinghua.ma [9]	0.1033	0.5571	0.1743	0.3154
biotext1trge	u.cberkeley.hearst [14]	0.0831	0.7	0.1486	0.3139
GUCply1260	german.u.cairo [18]	0.1444	0.4333	0.2167	0.305
OHSUSVMJ20	ohsu.hersh [12]	0.2309	0.3524	0.279	0.2937
biotext2trge	u.cberkeley.hearst [14]	0.095	0.5548	0.1622	0.2905
THIRcat03	tsinghua.ma [9]	0.0914	0.55	0.1567	0.2765
THIRcat05	tsinghua.ma [9]	0.1082	0.4167	0.1718	0.245
biotext3trge	u.cberkeley.hearst [14]	0.1096	0.4024	0.1723	0.2389
nusbird2004a	mlg.nus [33]	0.1373	0.3357	0.1949	0.2302
nusbird2004d	mlg.nus [33]	0.1349	0.2881	0.1838	0.1957
nusbird2004b	mlg.nus [33]	0.1163	0.3	0.1677	0.1861
eres2	u.edinburgh.sinclair [32]	0.1647	0.231	0.1923	0.1724
biotext4trge	u.cberkeley.hearst [14]	0.1271	0.2571	0.1701	0.1688
emet2	u.edinburgh.sinclair [32]	0.1847	0.2071	0.1953	0.1614
epub2	u.edinburgh.sinclair [32]	0.1729	0.2095	0.1895	0.1594
nusbird2004e	mlg.nus [33]	0.136	0.231	0.1712	0.1576
geneteam3	u.hospital.geneva [35]	0.1829	0.1833	0.1831	0.1424
edis2	u.edinburgh.sinclair [32]	0.1602	0.1857	0.172	0.137
wdtriage1	indiana.u.yang [27]	0.202	0.1476	0.1706	0.1185
eint2	u.edinburgh.sinclair [32]	0.1538	0.1619	0.1578	0.1174
NTU3v3N1c2	ntu.chen [34]	0.1553	0.1357	0.1449	0.0988
geneteam1	u.hospital.geneva [35]	0.1333	0.1333	0.1333	0.09
geneteam2	u.hospital.geneva [35]	0.1333	0.1333	0.1333	0.09
biotext5trge	u.cberkeley.hearst [14]	0.1192	0.1214	0.1203	0.0765
TRICSUSM	u.sanmarcos [31]	0.0792	0.1762	0.1093	0.0738
IBMIRLver1	ibm.india (none)	0.2053	0.0738	0.1086	0.0595
EMCTNOT1	tno.kraaij [19]	0.2	0.0143	0.0267	0.0114
Mean		0.1381	0.5194	0.1946	0.3303
MeSH *Mice*	rutgers.dayanik [16]	0.1502	0.8929	0.2572	0.6404

Table 7 - Data file contents and counts for annotation hierarchy subtasks.

File contents	Training data count	Test data count
Documents - PMIDs	504	378
Genes - Gene symbol, MGI identifier, and gene name for all used	1294	777
Document gene pairs - PMID-gene pairs	1418	877
Positive examples - PMIDs	178	149
Positive examples - PMID-gene pairs	346	295
Positive examples - PMID-gene-domain tuples	589	495
Positive examples - PMID-gene-domain-evidence tuples	640	522
Positive examples - all PMID-gene-GO-evidence tuples	872	693
Negative examples - PMIDs	326	229
Negative examples - PMID-gene pairs	1072	582

Table 8 - Annotation hierarchy subtask, sorted by F-score.

Run	Group (reference)	Precision	Recall	F-score
lgcad1	indiana.u.seki [13]	0.4415	0.7697	0.5611
lgcad2	indiana.u.seki [13]	0.4275	0.7859	0.5537
wiscWRT	u.wisconsin [17]	0.4386	0.6202	0.5138
wiscWT	u.wisconsin [17]	0.4218	0.6263	0.5041
dimacsAg3mh	rutgers.dayanik [16]	0.5344	0.4545	0.4913
NLMA1	nlm.umd.ul [15]	0.4306	0.5515	0.4836
wiscWR	u.wisconsin [17]	0.4255	0.5596	0.4834
NLMA2	nlm.umd.ul [15]	0.427	0.5374	0.4758
wiscW	u.wisconsin [17]	0.3935	0.5596	0.4621
KoikeyaHi1	u.tokyo (none)	0.3178	0.7293	0.4427
iowarun3	u.iowa [23]	0.3207	0.6	0.418
iowarun1	u.iowa [23]	0.3371	0.5434	0.4161
iowarun2	u.iowa [23]	0.3812	0.4505	0.413
BIOTEXT22	u.cberkeley.hearst [14]	0.2708	0.796	0.4041
BIOTEXT21	u.cberkeley.hearst [14]	0.2658	0.8141	0.4008
dimacsAl3w	rutgers.dayanik [16]	0.5015	0.3273	0.3961
GUCsvm0	german.u.cairo [18]	0.2372	0.7414	0.3595
GUCir50	german.u.cairo [18]	0.2303	0.8081	0.3584
geneteamA5	u.hospital.geneva [35]	0.2274	0.7859	0.3527
GUCir30	german.u.cairo [18]	0.2212	0.8404	0.3502
geneteamA4	u.hospital.geneva [35]	0.209	0.9354	0.3417
BIOTEXT24	u.cberkeley.hearst [14]	0.4452	0.2707	0.3367
GUCsvm5	german.u.cairo [18]	0.2052	0.9354	0.3366
cuhkrun3	chinese.u.hongkong (none)	0.4174	0.2808	0.3357
geneteamA2	u.hospital.geneva [35]	0.2025	0.9535	0.334
dimacsAabsw1	rutgers.dayanik [16]	0.5979	0.2283	0.3304
BIOTEXT23	u.cberkeley.hearst [14]	0.4437	0.2626	0.3299
geneteamA1	u.hospital.geneva [35]	0.1948	0.9778	0.3248
geneteamA3	u.hospital.geneva [35]	0.1938	0.9798	0.3235
GUCbase	german.u.cairo [18]	0.1881	1	0.3167
BIOTEXT25	u.cberkeley.hearst [14]	0.4181	0.2525	0.3149
cuhkrun2	chinese.u.hongkong (none)	0.4385	0.2303	0.302
cuhkrun1	chinese.u.hongkong (none)	0.4431	0.2283	0.3013
dimacsAp5w5	rutgers.dayanik [16]	0.5424	0.1939	0.2857
dimacsAw20w5	rutgers.dayanik [16]	0.6014	0.1677	0.2622
iowarun4	u.iowa [23]	0.1692	0.1333	0.1492
Mean		0.3600	0.5814	0.3824

Table 9 - Annotation hierarchy plus evidence code subtask, sorted by F-score.

Tag	Group (reference)	Precision	Recall	F-score
lgcab2	indiana.u.seki [13]	0.3238	0.6073	0.4224
lgcab1	indiana.u.seki [13]	0.3413	0.4923	0.4031
KoikeyaHiev1	u.tokyo (none)	0.2025	0.4406	0.2774
Mean		0.2892	0.5134	0.3676

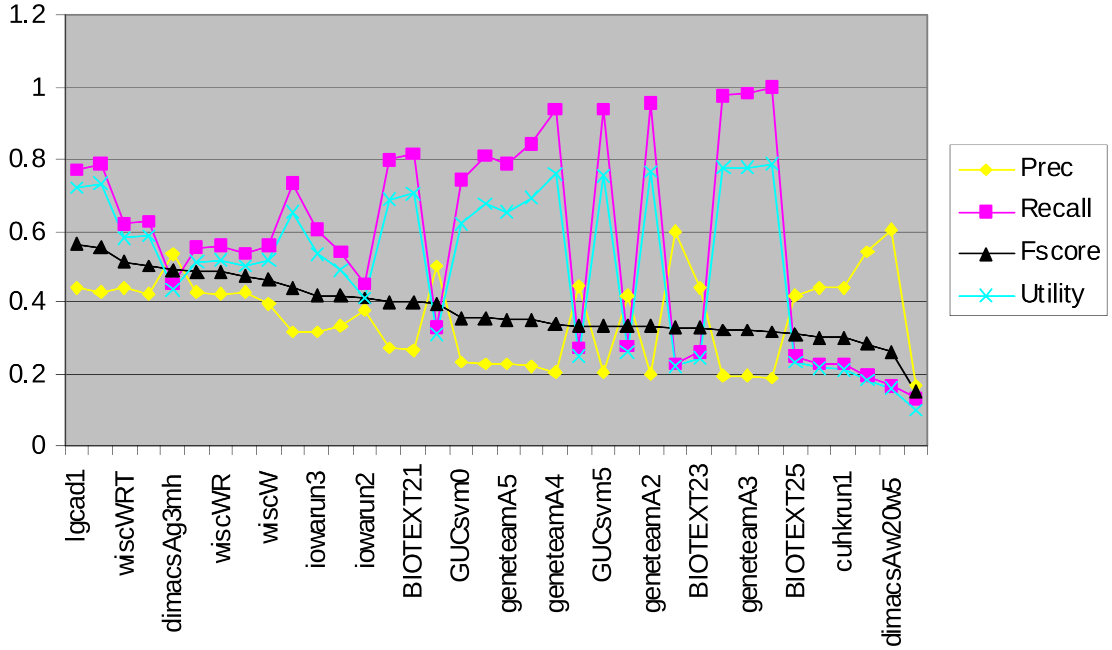

Figure 10 - Annotation hierarchy subtask results sorted by F-score.

In the annotation hierarchy subtask, the runs varied widely in recall and precision. The best runs, i.e., those with the highest F-scores, had medium levels of recall and precision. The top run came from Indiana University and used a variety of approaches, including a k-nearest neighbor model, mapping terms to MeSH, using keyword and glossary fields of documents, and recognizing gene names [13]. Further post-submission runs raised their F-score to 0.639. Across a number of groups, benefit was found from matching gene names appropriately. University of Wisconsin also found identifying gene names in sentences and modeling features in those sentences provided value [17].

5. Discussion

The TREC 2004 Genomics Track was very successful, with a great deal of enthusiastic participation. In all of the tasks, a diversity of approaches were used, resulting in wide variation across the results. Trying to discern the relative value of them is challenging, since few groups performed parameterized experiments or used common baselines.

In the ad hoc retrieval task, the best approaches employed techniques known to be effective in non-biomedical TREC tasks. These included Okapi weighting, blind relevance feedback, and language modeling. However, some domain-specific approaches appeared to be beneficial, such as expanding queries with synonyms from controlled vocabularies that are widely available. There also appeared to be some benefit for boosting parts of the queries. However, it was also easy for many groups to do detrimental things, as evidenced by the OHSU run of a TF*IDF system "out of the box" that scored well above the median.

The triage subtask was limited by the fact that using the MeSH term *Mice* assigned by the MEDLINE indexers was a better predictor of the MGI triage decision than anything else, including the complex feature extraction and machine learning algorithms of many participating groups. Some expressed concern that MGI might give preference to basing annotation decisions on maximizing coverage of genes instead of exhaustively cataloging the literature, something that would be useful for users of its system but compromise the value of its data in tasks like automated article triage. We were assured by the MGI director (J. Blake, personal communication) that the initial triage decision for an article was made independent of the prior coverage of gene, even though priority decisions made later in the pipeline did take coverage into account. As such, the triage decision upon which our data were based was sound from the standpoint of document classification. The annotation decision was also not effected by this since the positive and negative are not exhaustive (and do not need to be) for this subtask.

Another concern about the MGI data was whether the snapshot obtained in mid-2004 was significantly updated by the time the track was completed. This was analyzed in early 2005, and it was indeed found that the number of PMIDs in the triage subtask had increased in size by about 10%, with a very small number now negatively triaged. While this change is unlikely to have major impact on results, an updated data set will be released in early 2005.

But the remaining question for the triage subtask is why systems were unable to outperform the MeSH term *Mice*. It should be noted that this term was far from perfect, achieving a recall of 89% but a precision of only 15%. So why cannot more elaborate systems outperform this? There are a variety of explanations:

- MGI data is problematic - while MGI does some internal quality checking, they do not carry it out at the level that research groups would, e.g., with kappa scores
- Our algorithms and systems are imperfect - we do not know or there do not exist better predictive features
- Our metrics may be problematic - is the factor = 20 in the utility formula appropriate?

We believe that the triage subtask data represents an important task (i.e., document triage is valuable in a variety of biomedical settings, such as discerning the best evidence in clinical studies) and that these data provide the substrate for work to continue in this area.

The annotation hierarchy task had lower participation, and the value of picking the correct hierarchy is unclear. However, there would be great value to systems that could perform automated GO annotation, even though the task is very challenging [2]. These results demonstrated a value identifying gene names and other controlled vocabulary terms in documents for this task.

The TREC Genomics Track will be continuing in 2005. In addition, the data for the 2004 track will be released to the general community for continued experimentation. The categorization task data will be updated before its release, and both the old and new data will be made available. We hope that all of this will continue to facilitate in IR in the genomics domain.

Acknowledgements

The TREC 2004 Genomics Track was supported by NSF Grant ITR-0325160. The track also appreciates the help of Ellen Voorhees and NIST.

References

1. Mobasheri A, et al., *Post-genomic applications of tissue microarrays: basic research, prognostic oncology, clinical genomics and drug discovery.* Histology and Histopathology, 2004. 19: 325-335.
2. Hirschman L, et al., *Accomplishments and challenges in literature data mining for biology.* Bioinformatics, 2002. 18: 1553-1561.
3. Anonymous, *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Research, 2004. 32: D258-D261.
4. Hersh WR and Bhupatiraju RT. *TREC genomics track overview*. *The Twelfth Text Retrieval Conference: TREC 2003*. 2003. Gaithersburg, MD: National Institute of Standards and Technology. 14-23. http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.
5. Kramer MS and Feinstein AR, *Clinical biostatistics: LIV. The biostatistics of concordance.* Clinical Pharmacology and Therapeutics, 1981. 29: 111-123.
6. Hersh WR, et al. *OHSUMED: an interactive retrieval evaluation and new large test collection for research*. *Proceedings of the 17th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval. 1994. Dublin, Ireland: Springer-Verlag. 192-201.

7. Fujita S. *Revisiting again document length hypotheses - TREC 2004 Genomics Track experiments at Patolis*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/patolis.geo.pdf.
8. Buttcher S, Clarke CLA, and Cormack GV. *Domain-specific synonym expansion and validation for biomedical information retrieval (MultiText experiments for TREC 2004)*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uwaterloo-clarke.geo.pdf.
9. Li J, et al. *THUIR at TREC 2004: Genomics Track*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/tsinghua-ma.geo.pdf.
10. Carpenter B. *Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/alias-i.geo.pdf.
11. Pirkola A. *TREC 2004 Genomics Track experiments at UTA: the effects of primary keys, bigram phrases and query expansion on retrieval performance*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/utampere.geo.pdf.
12. Cohen AM, Bhuptiraju RT, and Hersh W. *Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/ohsu-hersh.geo.pdf.
13. Seki K, et al. *TREC 2004 Genomics Track experiments at IUB*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/indianau-seki.geo.pdf.
14. Nakov PI, et al. *BioText team experiments for the TREC 2004 Genomics Track*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/ucal-berkeley.geo.pdf.
15. Aronson AR, et al. *Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/nlm-umd-ul.geo.pdf.
16. Dayanik A, et al. *DIMACS at the TREC 2004 Genomics Track*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/rutgers-dayanik.geo.pdf.
17. Settles B and Craven M. *Exploiting zone information, syntactic rules, and informative terms in Gene Ontology annotation of biomedical documents*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uwisconsin.geo.pdf.
18. Darwish K and Madkour A. *The GUC goes to TREC 2004: using whole or partial documents for retrieval and classification in the Genomics Track*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/german.u.geo.pdf.
19. Kraaij W, et al. *MeSH based feedback, concept recognition and stacked classification for curation tasks*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/tno-emc.geo.pdf.
20. Crangle C, et al. *Concept extraction and synonymy management for biomedical information retrieval*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of

Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/converspeech.geo.pdf.

21. Billerbeck B, et al. *RMIT University at TREC 2004*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/rmit.tera.geo.pdf.
22. Tong RM. *Information needs and automatic queries*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/tarragon.tong.geo.pdf.
23. Eichmann D, et al. *Novelty, question answering and genomics: the University of Iowa response*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uiowa.novelty.qa.geo.pdf.
24. Bacchin M and Melucci M. *Expanding queries using stems and symbols*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/upadova.geo.pdf.
25. Ruiz ME, Srikanth M, and Srihari R. *UB at TREC 13: Genomics Track*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/stateuny-buffalo.geo.pdf.
26. Huang X, et al. *York University at TREC 2004: HARD and Genomics Tracks*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/yorku.hard.geo.pdf.
27. Yang K, et al. *WIDIT in TREC 2004 Genomics, Hard, Robust and Web Tracks*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/indianau.geo.hard.robust.web.pdf.
28. Blott S, et al. *Experiments in terabyte searching, genomic retrieval and novelty detection for TREC 2004*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/dcu.tera.geo.novelty.pdf.
29. Guo Y, Harkema H, and Gaizauskas R. *Sheffield University and the TREC 2004 Genomics Track: query expansion using synonymous terms*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/usheffield.geo.pdf.
30. Tomiyama T, et al. *Meiji University Web, Novelty and Genomic Track experiments*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/meijiu.web.novelty.geo.pdf.
31. Guillen R. *Categorization of genomics text based on decision rules*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/cal-state-sanmarcos.geo.pdf.
32. Sinclair G and Webber B. *TREC Genomics 2004*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uedinburgh-sinclair.geo.pdf.
33. Zhang D and Lee WS. *Experience of using SVM for the triage task in TREC 2004 Genomics Track*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/natusing.zhang.geo.pdf.
34. Lee C, Hou WJ, and Chen HH. *Identifying relevant full-text articles for GO annotation without MeSH terms*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/ntu.geo.pdf.
35. Ruch P, et al. *Report on the TREC 2004 experiment: Genomics Track*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uhosp-geneva.geo.pdf.