TREC 2005 Genomics Track Overview

William Hersh[1], Aaron Cohen[1], Jianji Yang[1], Ravi Teja Bhupatiraju[1],
Phoebe Roberts[2], Marti Hearst[3]
[1]Oregon Health & Science University, Portland, OR, USA
[2]Biogen Idec Corp., Boston, MA, USA
[3]University of California, Berkeley, CA, USA

The TREC 2005 Genomics Track featured two tasks, an ad hoc retrieval task and four subtasks in text categorization. The ad hoc retrieval task utilized a 10-year, 4.5-million document subset of the MEDLINE bibliographic database, with 50 topics conforming to five generic topic types. The categorization task used a full-text document collection with training and test sets consisting of about 6,000 biomedical journal articles each. Participants aimed to triage the documents into categories representing data resources in the Mouse Genome Informatics database, with performance assessed via a utility measure.

1. Introduction

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 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. There are many reasons why a focus on this domain is important. New advances in biotechnologies have changed the face of biological research, particularly "high-throughput" techniques such as gene microarrays [1]. These techniques 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 and factors that require further investigation. Furthermore, the literature itself becomes a source of "experiments" as researchers turn to it to search for knowledge that in turn 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, 3].

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 the 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 towards structured 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.

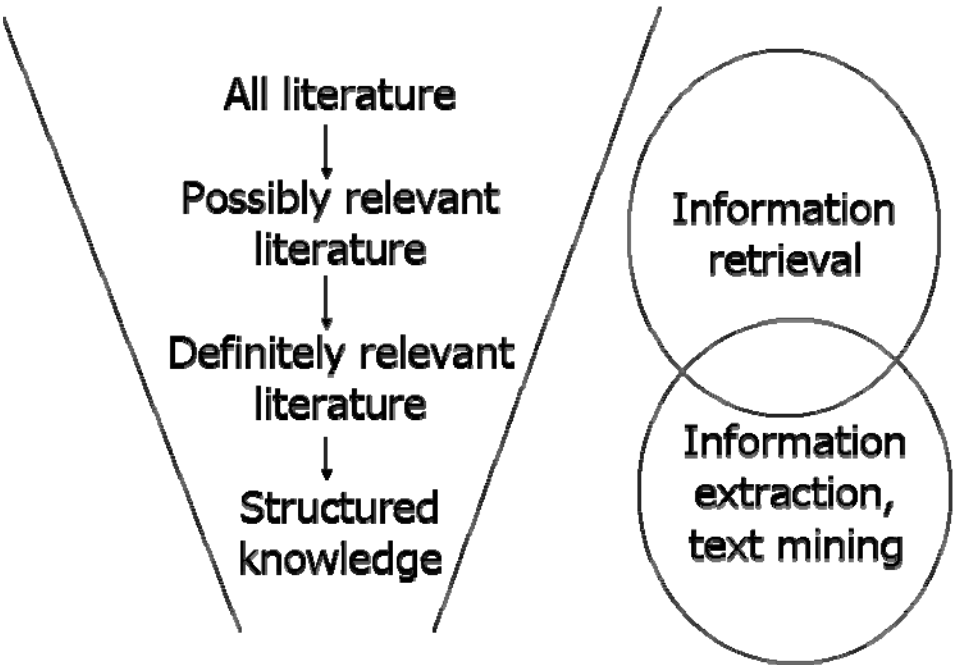

Figure 1 - The funneling of scientific literature and related information retrieval and extraction disciplines.

TREC 2005 marks the third offering of the Genomics Track. The first of the track, 2003, was limited by lack of resources to perform relevance judgments and other tasks, so the track had to use "pseudojudgments" culled from data created for other purposes [4]. In 2004, however, the track obtained a five-year grant from the U.S. National Science Foundation (NSF), which provided resources for building test collections and other data sources. The 2004 track featured an ad hoc retrieval task [5] and three subtasks in text categorization [6].

For 2005, the track built on the success of 2004 by using the same underlying document collections on new topics for ad hoc retrieval and refinement of the text categorization tasks. Similar to the 2004 track, the track attracted the largest number of participating groups of any in TREC. In 2005, 32 groups submitted 59 runs to the ad hoc retrieval task, while 19 groups submitted 192 runs to the categorization subtasks. A total of 41 different groups participated, with 10 groups participating in both tasks, 22 participating only in the ad hoc retrieval task, and 9 participating in just the categorization tasks, making it the largest track in TREC 2005.

The remainder of this paper covers the tasks, methods, and results of the two tasks separately, followed by discussion of future directions.

2. Ad Hoc Task

2.1 Task

The ad hoc retrieval task modeled the situation of a user with an information need using an information retrieval system to access the biomedical scientific literature. The document collection was based on a large subset of the MEDLINE bibliographic database. It should be

noted that although we are in an era of readily available full-text journals (usually requiring a subscription), many users of the biomedical literature enter through searching MEDLINE. As such, there are still strong motivations to improve the effectiveness of searching MEDLINE.

2.2 Documents

The document collection for the 2005 ad hoc retrieval task was the same 10-year MEDLINE subset using for the 2004 track. One goal we have is to produce a number of topic and relevance judgment collections that use this same document collection to make retrieval experimentation easier (so people do not have to load different collections into their systems). Additional uses of this subset have already appeared [7]. MEDLINE can be searched by anyone in the world using the PubMed system of the National Library of Medicine (NLM), which maintains both MEDLINE and PubMed. The full MEDLINE database contains over 14 million references dating back to 1966 and is updated on a daily basis.

The subset of MEDLINE for the TREC 2005 Genomics 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, which is about one third of the full MEDLINE database. The data included all of the PubMed fields identified in the MEDLINE Baseline record. Descriptions of the various fields of MEDLINE are available at: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#MEDLINEDisplayFormat

The MEDLINE subset was provided in the "MEDLINE" format, consisting of ASCII text with fields indicated and delimited by 2-4 character abbreviations. The size of the file uncompressed was 9,587,370,116 bytes. An XML version of MEDLINE subset was also available. It should also be noted that not all MEDLINE records have abstracts, usually because the article itself does not have an abstract. In general, about 75% of MEDLINE records have abstracts. In our subset, there were 1,209,243 (26.3%) records without abstracts.

2.3 Topics

As with 2004, we collected information needs from real biologists. However, instead of soliciting free-form biomedical questions, we developed a set of six generic topic templates (GTTs) derived from an analysis of the topics from the 2004 track and other known biologist information needs (Table 1). GTTs consist of semantic types, such as genes or diseases, placed in the context of commonly queried biomedical questions, and semantic types are often present in more than one GTT. After we developed the GTTs, 11 people interviewed 25 biologists to obtain ten or more specific information needs that conformed to each GTT. One GTT did not model a commonly researched problem, and was dropped from the study. The topics did not have to fit precisely into the GTTs, but had to come close, i.e., have all the required semantic types. We then had other people search on the topics to make sure there was some, but not too much, relevant information in MEDLINE.). Ten information needs for each GTT were selected for inclusion in the 2005 track to total fifty topics.

In order to get participating groups started with the topics, and in order for them not to "spoil"

their automatic status of their official runs by working with the official topics, we developed 10 sample topics, consisting of two topics from each GTT. These learning topics had a MEDLINE search and relevance judgments of the output that we made available to participants. Table 1 also gives an example topic for each GTT that comes from the sample topics.

2.4 Relevance judgments

Relevance judgments were done using the conventional pooling method of TREC. Based on estimation of relevance judgment resources, the top 60 documents for each topic from all official runs were used. This gave an average pool size of 821 documents with a range of 290 to 1356. These pools were then provided to the relevance judges, who consisted of five individuals with varying expertise in biology. The relevance judges were instructed in the following manner for each GTT:

- 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.

The articles had to describe a specific gene, disease, impact, mutation, etc. and not just the concept in general.

Table 1 - Generic topic types and example sample topics. The semantic types in each GTT are underlined.

Generic Topic Type	**Topic Range**	**Example Sample Topic**
Find articles describing standard methods or protocols for doing some sort of experiment or procedure	100-109	Method or protocol: GST fusion protein expression in Sf9 insect cells
Find articles describing the role of a gene involved in a given disease	110-119	Gene: DRD4 Disease: Alcoholism
Find articles describing the role of a gene in a specific biological process	120-129	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	130-139	Genes: HMG and HMGB1 Disease: Hepatitis
Find articles describing one or more mutations of a given gene and its biological impact	140-149	Gene with mutation: Ret Biological impact: Thyroid function

Relevance judges were asked to rate documents as definitely, possibly, or not relevant. As in 2004, articles that were rated definitely or possibly relevant were considered relevant for use in the binary recall and precision-related measures of retrieval performance. Relevance judgments were performed by individuals with varying levels of expertise in biology (from an undergraduate student to a PhD researcher). For 10 of the topics, judgments were performed in duplicate to allow interobserver reliability measurement using the kappa statistic.

2.5 Measures and statistical analysis

Retrieval performance was measured with the "usual" TREC ad hoc measures of mean average precision (MAP), binary preference (B-Pref) [8], precision at the point of the number of relevant documents retrieved (R-Prec), and precision at varying numbers of documents retrieved (e.g., 5, 10, 30, etc. documents up to 1,000). These measures were calculated using version 8.0 of trec_eval developed by Chris Buckley (Sabir Research).

Research groups submitted their runs through the TREC Web site in the usual manner. They were 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

They were also required to provide a brief system description.

Statistical analysis of the above measures was performed using SPSS (version 12.0). Repeated measure analysis of variance (ANOVA) with posthoc tests using Sidak adjustments were performed on the above variables. In addition, descriptive analysis of MAP was also done to study the spread of the data.

2.6 Results

A total of 32 groups submitted 58 runs. Table 2 shows the results of relevance judging for each topic, listing the pool size sent to a given assessor plus their distribution of relevance assessments. The combined number and percentage of documents rated definitely and possibly relevant are also listed, since these were considered relevant from the standpoint of official results. Six topics had no definitely relevant documents. One topic had no definitely or possibly relevant documents and was dropped from the calculation of official results.

Table 2 - Relevant documents per topic. Topic 135 had no relevant documents and was eliminated from the results. Documents that were definitely or possibly relevant were considered to be relevant for the purposes of official TREC results.

Topic	Pool Size	Definitely Relevant	Possibly Relevant	Not Relevant	Definitely + Possibly (TREC) Relevant	% TREC Relevant
100	704	22	52	630	74	10.5%
101	651	2	18	631	20	3.1%
102	1164	5	5	1154	10	0.9%
103	701	6	19	676	25	3.6%
104	629	0	4	625	4	0.6%
105	1133	4	85	1044	89	7.9%
106	1230	44	125	1061	169	13.7%
107	484	76	114	294	190	39.3%
108	1092	76	127	889	203	18.6%
109	389	165	14	210	179	46.0%
110	934	4	12	918	16	1.7%
111	675	109	93	473	202	29.9%
112	872	4	7	861	11	1.3%
113	1356	10	4	1342	14	1.0%
114	754	210	169	375	379	50.3%
115	1350	3	12	1335	15	1.1%
116	1265	58	28	1179	86	6.8%
117	1094	527	182	385	709	64.8%
118	938	20	12	906	32	3.4%
119	589	42	19	528	61	10.4%
120	527	223	122	182	345	65.5%
121	422	17	25	380	42	10.0%
122	871	19	37	815	56	6.4%
123	1029	5	32	992	37	3.6%
124	752	8	53	691	61	8.1%
125	1202	3	8	1191	11	0.9%
126	1320	190	117	1013	307	23.3%
127	841	1	3	837	4	0.5%
128	954	21	53	880	74	7.8%
129	987	16	22	949	38	3.9%
130	813	9	23	781	32	3.9%
131	431	2	40	389	42	9.7%
132	531	3	27	501	30	5.6%
133	523	0	5	518	5	1.0%
134	732	2	9	721	11	1.5%
135	1057	0	0	1057	0	0.0%
136	853	0	3	850	3	0.4%
137	1129	12	39	1078	51	4.5%
138	501	6	6	489	12	2.4%
139	380	15	20	345	35	9.2%
140	395	14	15	366	29	7.3%
141	520	34	47	439	81	15.6%
142	528	151	120	257	271	51.3%
143	902	0	4	898	4	0.4%
144	1212	1	1	1210	2	0.2%
145	288	10	22	256	32	11.1%
146	825	370	67	388	437	53.0%
147	659	0	10	649	10	1.5%
148	536	0	11	525	11	2.1%
149	1294	6	17	1271	23	1.8%
Avg	820.4	50.5	41.2	728.7	91.7	12.5%

Table 3 - Overlap of duplicate judgments for kappa statistic.

	Duplicate judge - Relevant	**Duplicate judge - Not Relevant**	**Total**
Original judge - Relevant	1100	629	1729
Original judge - Not Relevant	546	8204	8750
Total	1646	8833	10479

In order to assess the consistency of relevance judgments, we had judgments of ten topics performed in duplicate. (For three topics, we actually had judgments performed in triplicate; one of these was the topic that had no relevant documents.) The judgments from the original judge who did the assessing was used as the "official" judgment. Table 3 shows the consistency of the judgments from the original and duplicating judge. The kappa score for inter-judge agreement was 0.585, indicating a "moderate" level of agreement and comparable to the 2004 Genomics Track.

The overall results are shown in Table 4, sorted by MAP. The top-ranking run came from York University. The top-ranking run was a manual run, but this group also had the top-ranking automatic run. The top-ranking interactive run was somewhat further down the list, although this group had an automatic run that performed better. The statistical analysis of the runs showed overall statistical significance for all of the measures. Pair-wise comparison of MAP for the 58 runs showed that significant difference from the top run was obtained at run uta05i. At the other end, significant difference from the lowest run was reached by run genome2. Figure 2 shows the MAP results with 95% confidence intervals, while Figure 3 shows all of the statistics from Table 4, sorted by each run's MAP.

We also assessed the results by topic. Table 5 shows the various measures for each topic, while Figure 4 shows the same data graphically with confidence intervals. The spread of MAP showed a wide variation among the 49 topics. Topic 136 had the lowest variance (<0.001) with range of 0-0.0287. On the other hand, topic 119 showed the highest variance (0.060), with range of 0.0144-0.8289. Topic 121 received the highest mean MAP at 0.620, while topic 143 had the lowest at 0.003. Figure 5 compares the number of relevant documents with MAP for each topic.

In addition, we grouped the results by GTT, as shown in Table 6. The GTT of information describing the role of a gene in a disease achieved the highest MAP, while the gene interactions and gene mutations achieved the best B-Pref. However, the differences among all of the GTTs were modest.

Table 4 - Run results by run name, type (manual, automatic, or interactive), and performance measures.

Run	Group	Type	MAP	R-Prec	B-pref	P10	P100	P1000
york05gm1 [9]	yorku.huang	m	0.302	0.3212	0.3155	0.4551	0.2543	0.0748
york05ga1 [9]	yorku.huang	a	0.2888	0.3118	0.3061	0.4592	0.2557	0.0721
ibmadz05us [10]	ibm.zhang	a	0.2883	0.3091	0.3026	0.4735	0.2643	0.0766
ibmadz05bs [10]	ibm.zhang	a	0.2859	0.3061	0.2987	0.4694	0.2606	0.0761
uwmtEg05	uwaterloo.clarke	a	0.258	0.2853	0.2781	0.4143	0.2292	0.0718
UIUCgAuto [11]	uiuc.zhai	a	0.2577	0.2688	0.2708	0.4122	0.231	0.0709
UIUCgInt [11]	uiuc.zhai	i	0.2487	0.2627	0.267	0.4224	0.2355	0.0694
NLMfusionA [12]	nlm-umd.aronson	a	0.2479	0.2767	0.2675	0.402	0.2378	0.0688
iasl1 [13]	academia.sinica.tsai	a	0.2453	0.2708	0.265	0.398	0.2292	0.0698
NLMfusionB [12]	nlm-umd.aronson	a	0.2453	0.2666	0.2541	0.4082	0.2339	0.0693
UniNeHug2 [14]	uneuchatel.savoy	a	0.2439	0.2582	0.264	0.398	0.2308	0.0712
UniGe2 [15]	u.geneva	a	0.2396	0.2705	0.2608	0.3878	0.2361	0.0711
i2r1 [16]	iir.yu	a	0.2391	0.2629	0.2716	0.3898	0.231	0.0668
uta05a [17]	utampere.pirkola	a	0.2385	0.2638	0.2546	0.4163	0.2255	0.0678
i2r2 [16]	iir.yu	a	0.2375	0.2622	0.272	0.3878	0.2296	0.067
UniNeHug2c [14]	uneuchatel.savoy	a	0.2375	0.2662	0.2589	0.3878	0.239	0.0725
uwmtEg05fb	uwaterloo.clarke	a	0.2359	0.2573	0.2552	0.3878	0.2257	0.0712
DUTAdHoc2 [18]	dalianu.yang	m	0.2349	0.2678	0.2725	0.3939	0.2206	0.0648
THUIRgen1S [19]	tsinghua.ma	a	0.2349	0.2663	0.2568	0.4224	0.2214	0.0622
tnog10 [20]	tno.erasmus.kraaij	a	0.2346	0.2607	0.2564	0.3857	0.2227	0.0668
DUTAdHoc1 [18]	dalianu.yang	m	0.2344	0.2718	0.2726	0.402	0.22	0.0645
tnog10p [20]	tno.erasmus.kraaij	a	0.2332	0.2506	0.2555	0.402	0.2173	0.0668
iasl2 [13]	academia.sinica.tsai	a	0.2315	0.2465	0.2487	0.3816	0.2276	0.07
UAmscombGeFb [21]	uamsterdam.aidteam	a	0.2314	0.2638	0.2592	0.4163	0.2271	0.0612
UBIgeneA [22]	suny-buffalo.ruiz	a	0.2262	0.2567	0.2542	0.3633	0.2122	0.0683
OHSUkey [23]	ohsu.hersh	a	0.2233	0.2569	0.2544	0.3735	0.2169	0.0632
NTUgah2 [24]	ntu.chen	a	0.2204	0.2562	0.2498	0.398	0.1996	0.0644
THUIRgen2P [19]	tsinghua.ma	a	0.2177	0.2519	0.2395	0.4143	0.2198	0.0695
NTUgah1 [24]	ntu.chen	a	0.2173	0.2558	0.2513	0.3918	0.1998	0.0615
UniGeNe [15]	u.geneva	a	0.215	0.2364	0.2347	0.3367	0.2237	0.0694
UAmscombGeMl [21]	uamsterdam.aidteam	a	0.2015	0.2325	0.232	0.3551	0.2094	0.0568
uta05i [17]	utampere.pirkola	i	0.198	0.2411	0.229	0.4082	0.2137	0.0547
PDnoSE [25]	upadova.bacchin	a	0.1937	0.2213	0.2183	0.3571	0.2006	0.063
iitprf011003 [26]	iit.urbain	a	0.1913	0.2142	0.2205	0.3612	0.2018	0.065
dcu1 [27]	dublincityu.gurrin	a	0.1851	0.2178	0.2129	0.3816	0.1851	0.0577
dcu2 [27]	dublincityu.gurrin	a	0.1844	0.2234	0.214	0.3959	0.1896	0.0599
SFUshi [28]	simon-fraseru.shi	m	0.1834	0.2072	0.2149	0.3429	0.1898	0.0608
OHSUall [23]	ohsu.hersh	a	0.183	0.2285	0.2221	0.3286	0.1965	0.0592
wim2 [29]	fudan.niu	a	0.1807	0.2006	0.2055	0.3	0.1794	0.057
genome1 [30]	csusm.guillen	a	0.1803	0.2174	0.211	0.3245	0.1749	0.0577
wim1 [29]	fudan.niu	a	0.1781	0.2094	0.2076	0.3347	0.181	0.0592
NCBITHQ [12]	nlm.wilbur	a	0.1777	0.214	0.2192	0.3041	0.1824	0.0526
NCBIMAN [12]	nlm.wilbur	m	0.1747	0.2081	0.2181	0.3122	0.182	0.0519
UICgen1 [31]	uillinois-chicago.liu	a	0.1738	0.2079	0.2046	0.3082	0.1941	0.0579
MARYGEN1 [32]	umaryland.oard	a	0.1729	0.1954	0.1898	0.3041	0.1439	0.0409
PDSESe02 [25]	upadova.bacchin	a	0.1646	0.1928	0.1928	0.3224	0.1904	0.0615
genome2 [30]	csusm.guillen	a	0.1642	0.1931	0.1928	0.298	0.1676	0.0565
UIowa05GN102 [33]	uiowa.eichmann	a	0.1303	0.1861	0.1693	0.2898	0.1671	0.0396
UMD01 [34]	umichigan-dearborn.murphey	a	0.1221	0.1541	0.1435	0.3224	0.1473	0.0321
UIowa05GN101 [33]	uiowa.eichmann	a	0.1095	0.1636	0.1414	0.2857	0.1571	0.026
CCP0 [35]	ucolorado.cohen	m	0.1078	0.1486	0.1311	0.2837	0.1439	0.0203
YAMAHASHI2	utokyo.takahashi	m	0.1022	0.1236	0.1276	0.2653	0.1312	0.0369
YAMAHASHI1	utokyo.takahashi	m	0.1003	0.1224	0.1248	0.2531	0.1267	0.0356
dpsearch2 [36]	datapark.zakharov	m	0.0861	0.1169	0.1034	0.2633	0.1231	0.0278
dpsearch1 [36]	datapark.zakharov	m	0.0827	0.1177	0.1017	0.2551	0.1182	0.0274
asubaral	arizonau.baral	m	0.0797	0.1079	0.0967	0.2714	0.1061	0.0142
CCP1 [35]	ucolorado.cohen	m	0.0554	0.0963	0.0775	0.1878	0.0951	0.0134
UMD02 [34]	umichigan-dearborn.murphey	a	0.0544	0.0703	0.0735	0.1755	0.0843	0.0166
Minimum			0.0544	0.0703	0.0735	0.1755	0.0843	0.0134
Mean			0.1968	0.2258	0.2218	0.3576	0.1976	0.0573
Maximum			0.302	0.3212	0.3155	0.4735	0.2643	0.0766

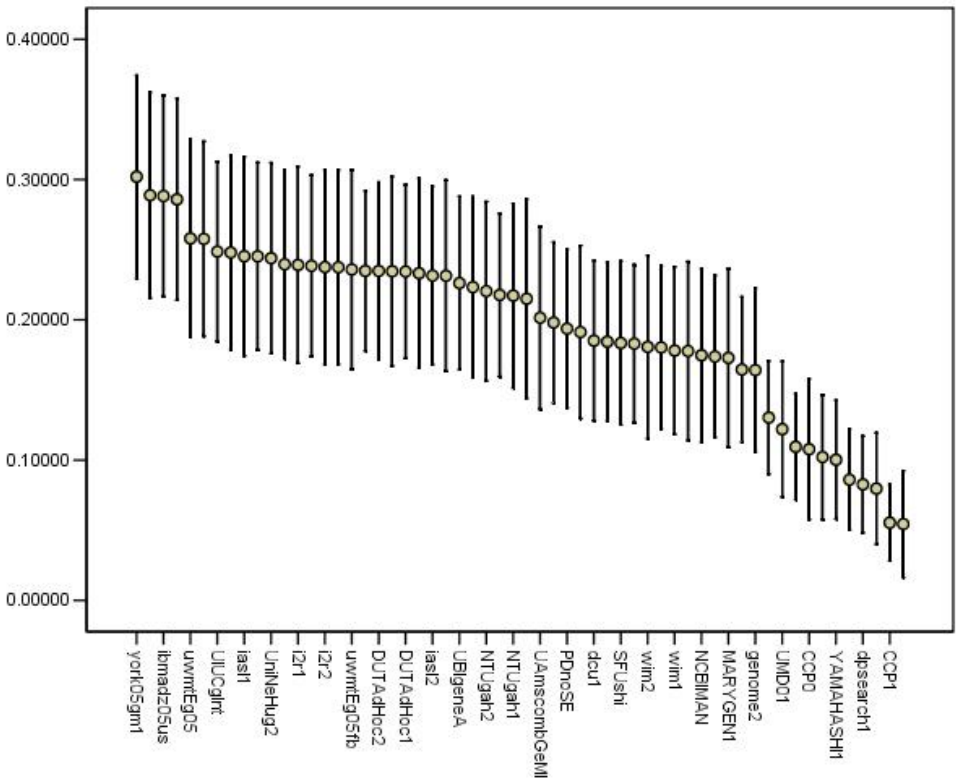

Figure 2 - Run results with 95% confidence intervals, sorted alphabetically.

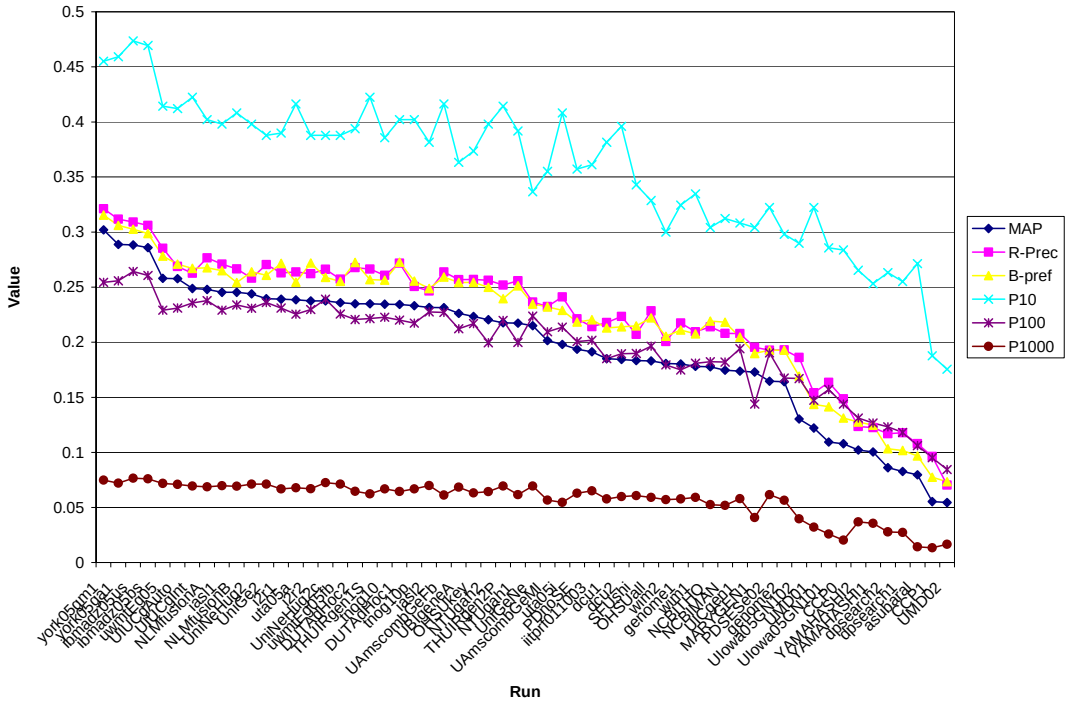

Figure 3 - Run results plotted graphically, sorted by MAP of each run.

Table 5 - Results by topic.

Topic	MAP	R-Prec	B-Pref	P10	P100	P1000
100	0.1691	0.2148	0.1616	0.3569	0.1916	0.0550
101	0.0454	0.0526	0.0285	0.0483	0.0516	0.0141
102	0.0110	0.0172	0.0100	0.0172	0.0091	0.0036
103	0.0603	0.0945	0.0570	0.0948	0.0602	0.0169
104	0.0694	0.0948	0.0582	0.0690	0.0124	0.0023
105	0.1102	0.1703	0.1461	0.4655	0.1586	0.0327
106	0.0625	0.1120	0.1231	0.3138	0.1433	0.0491
107	0.4184	0.4297	0.5289	0.9103	0.5934	0.1373
108	0.1224	0.1973	0.2206	0.4828	0.2788	0.0695
109	0.5347	0.5196	0.6512	0.9190	0.7066	0.1345
110	0.0137	0.0248	0.0154	0.0224	0.0128	0.0055
111	0.2192	0.2985	0.2926	0.3569	0.3140	0.1170
112	0.2508	0.3354	0.2754	0.3586	0.0481	0.0062
113	0.3124	0.3498	0.3164	0.3931	0.0822	0.0096
114	0.3876	0.4364	0.5505	0.8259	0.6697	0.2476
115	0.0378	0.0437	0.0340	0.0534	0.0193	0.0036
116	0.1103	0.1720	0.1456	0.2879	0.1636	0.0359
117	0.3796	0.4739	0.5126	0.8345	0.7409	0.4099
118	0.1343	0.1460	0.1369	0.3276	0.0634	0.0145
119	0.5140	0.5212	0.5075	0.8190	0.3462	0.0493
120	0.5769	0.5421	0.7217	0.9259	0.8091	0.2695
121	0.6205	0.6560	0.6394	0.7983	0.3040	0.0337
122	0.1423	0.2023	0.1590	0.3569	0.1510	0.0320
123	0.0375	0.0708	0.0474	0.1121	0.0493	0.0133
124	0.1519	0.2035	0.1693	0.5103	0.1505	0.0324
125	0.0772	0.0862	0.0708	0.0897	0.0209	0.0028
126	0.1313	0.2172	0.2388	0.3966	0.2979	0.1422
127	0.1015	0.1250	0.0862	0.0759	0.0155	0.0028
128	0.0921	0.1424	0.1062	0.3224	0.1247	0.0366
129	0.0864	0.1393	0.0939	0.1793	0.0984	0.0212
130	0.3390	0.3545	0.3346	0.6362	0.1388	0.0194
131	0.4436	0.4384	0.4230	0.5517	0.2790	0.0343
132	0.1048	0.1558	0.1115	0.2431	0.0966	0.0196
133	0.0328	0.0207	0.0172	0.0172	0.0140	0.0029
134	0.1687	0.1771	0.1582	0.1914	0.0364	0.0069
136	0.0032	0.0000	0.0000	0.0000	0.0019	0.0010
137	0.0676	0.1146	0.0767	0.1776	0.0848	0.0232
138	0.2196	0.2342	0.2029	0.2534	0.0552	0.0089
139	0.3600	0.3941	0.3488	0.5810	0.2052	0.0305
140	0.2700	0.3115	0.2423	0.3810	0.1843	0.0248
141	0.2381	0.2735	0.2053	0.3362	0.2598	0.0699
142	0.4416	0.4608	0.5911	0.8569	0.6409	0.2098
143	0.0031	0.0043	0.0011	0.0034	0.0021	0.0009
144	0.0734	0.0603	0.0431	0.0276	0.0053	0.0009
145	0.3363	0.3761	0.3238	0.5931	0.1852	0.0260
146	0.4808	0.4961	0.6325	0.8466	0.7212	0.3076
147	0.0087	0.0138	0.0057	0.0138	0.0091	0.0040
148	0.0411	0.0376	0.0144	0.0293	0.0407	0.0066
149	0.0286	0.0495	0.0304	0.0603	0.0347	0.0089

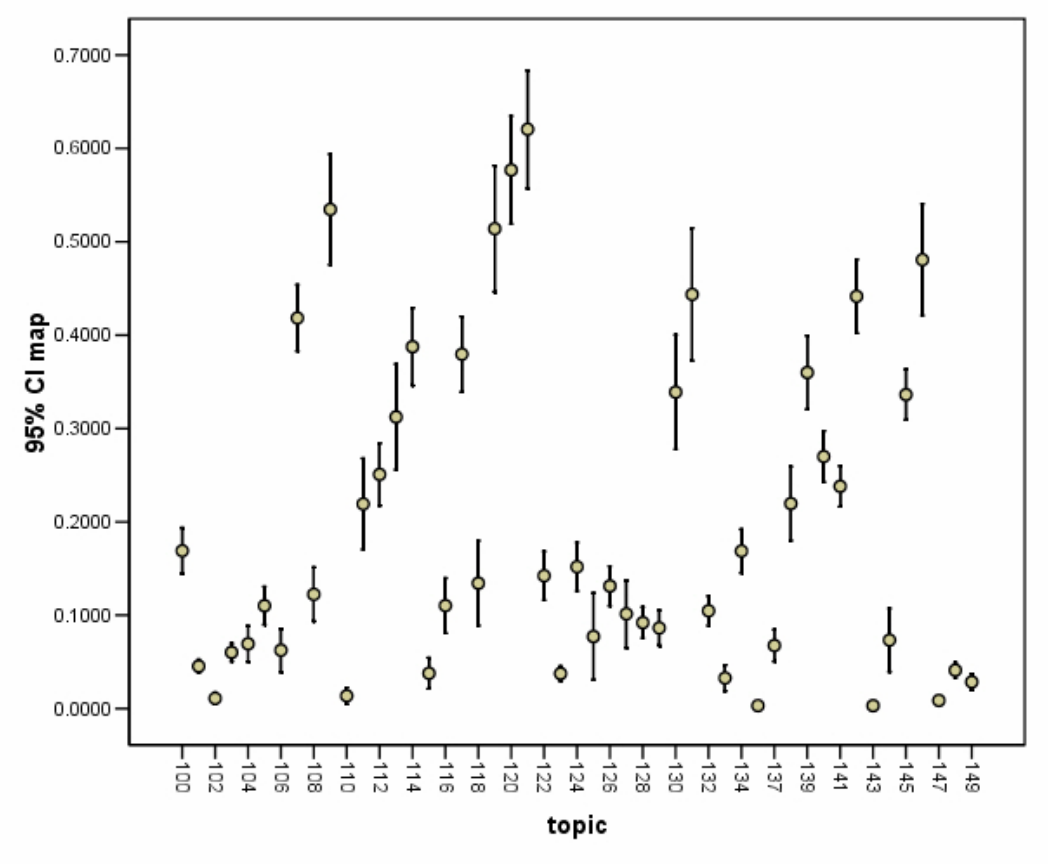

Figure 4 - Results by topic plotted graphically.

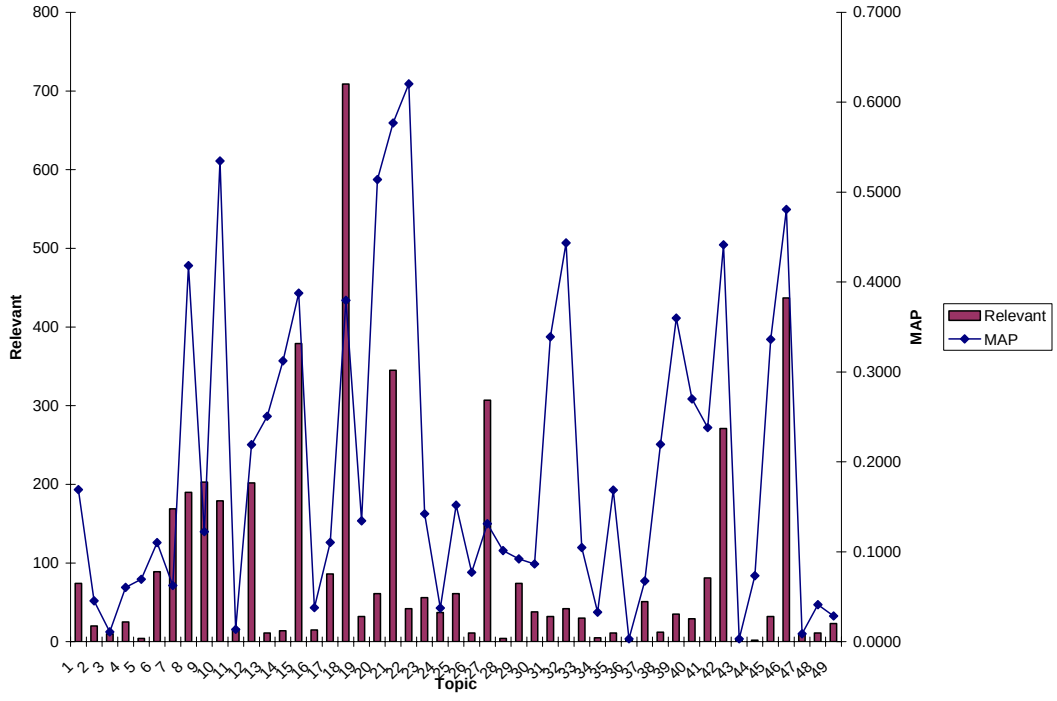

Figure 5 - Comparison of number of relevant documents and MAP for each topic.

Table 6 - Results by generic topic type.

Topics	GTT	MAP	R-Prec	B-Pref	P10	P100	P1000
100-109	Information describing standard methods or protocols for doing some sort of experiment or procedure	0.1603	0.1903	0.1985	0.3678	0.2206	0.0515
110-119	Information describing the role(s) of a gene involved in a disease	0.2360	0.2802	0.2787	0.4279	0.2460	0.0899
120-129	Information describing the role of a gene in a specific biological process	0.2018	0.2385	0.2333	0.3767	0.2021	0.0587
130-139	Information describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease	0.1932	0.2099	0.1859	0.2946	0.1013	0.0163
140-149	Information describing one or more mutations of a given gene and its biological impact or role	0.1922	0.2084	0.2090	0.3148	0.2083	0.0659

3. Categorization Task

3.1 Subtasks

The second task for the 2005 track was a full-text document categorization task. It was similar in part to the 2004 categorization task in using data from the Mouse Genome Informatics (MGI, http://www.informatics.jax.org/) system [37] and was a document triage task, where a decision is made on a per-document basis about whether or not to pass a document on for further expert review. It included a repeat of one subtask from last year, the triage of articles for GO annotation [38], and added triage of articles for three other major types of information collected and catalogued by MGI. These include articles about tumor biology [39], embryologic gene expression [40], and alleles of mutant phenotypes [41].

As such, the categorization task assessed how well systems can categorize documents in four separate categories. We used the same utility measure used last year but with different parameters (see below). We created an updated version of the cat_eval program that calculated the utility measure plus recall, precision, and the F score.

3.2 Documents

The documents for the 2005 categorization tasks consisted of the same full-text articles used in 2004. The articles came from three journals over two years, reflecting the full-text data we were able to obtain from Highwire Press: *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 available in SGML format based on Highwire's document type definition (DTD). Also the same as 2004, we designated articles published in 2002 as training data and those in 2003 as test data.

The documents for the tasks come from a subset of these articles that have the words “mouse” or “mice” or “murine” as described in the 2004 protocol. A crosswalk (look-up) table was provided that matches an identifier for each Highwire article (its file name) to its corresponding PubMed ID (PMID). Table 7 shows the total number of articles and the number in the subset the track used.

The training document collection was 150 megabytes in size compressed and 449 megabytes uncompressed. The test document collection was 140 megabytes compressed and 397 megabytes uncompressed. Many gene names have Greek or other non-English characters, which can present a problem for those attempting to recognize gene names in the text. The Highwire SGML appears to obey the rules posted on the NLM Web site with regards to these characters (http://www.ncbi.nlm.nih.gov/entrez/query/static/entities.html).

3.3 Data

The data for the triage decisions were provided by MGI. They were reformatted in a way to allow easy use by track participants and the cat_eval evaluation program.

3.4 Evaluation Measures

While we again used the utility measure as the primary evaluation measure, we used it in a slightly different way in 2005. This was because there were varying numbers of positive examples for the four different categorization tasks. The framework for evaluation in the categorization task is based on the possibilities in Table 8. The utility measure is often applied in text categorization research and was 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:
$U_{norm} = U_{raw} / U_{max}$

Table 7 - Distribution of documents in training and test sets.

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

Table 8 - Categories for utility measures.

	Relevant (classified)	**Not relevant (not classified)**	**Total**
Retrieved	True positive (TP)	False positive (FP)	All retrieved (AR)
Not retrieved	False negative (FN)	True negative (TN)	All not retrieved (ANR)
	All positive (AP)	All negative (AN)	

For a given test collection of documents to categorize, U_{raw} is calculated as follows:
$U_{raw} = (u_r * TP) + (u_{nr} * FP)$
where:

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

For our purposes, we assume that $u_{nr} = -1$ and solve for u_r assigning MGI's current practice of triaging everything a utility of 0.0:

$0.0 = u_r*AP - AN$
$u_r = AN/AP$

AP and AN are different for each task, as shown in Table 9. (The numbers for GO annotation are slightly different from the 2004 data. This is because additional articles have been triaged by MGI since we used that data last year.)

The u_r values for A and G are fairly close across the training and test collections, while they vary much more for E and especially T. We therefore established a u_r that was the average of that computed for the training and test collections, rounded to the nearest whole number. The resulting values for u_r for each subtask are shown in Table 10. In order to facilitate calculation of the modified version of the utility measure for the 2005 track, we updated the cat_eval program to version 2.0, which included a command-line parameter to set u_r. The training and test data were provided in four files, one for each category (i.e., A, E, G, and T). (The fact that three of those four corresponded to the four nucleotides in DNA was purely coincidental! We could not think of a good way to make a C from embryonic expression.)

Table 9 - Calculating u_r for subtasks.

Subtask	**Training**				**Test**			
	N	AP	AN	u_r	N	AP	AN	u_r
A (alelle)	5837	338	5499	16.27	6043	332	5711	17.20
E (expression)	5837	81	5756	71.06	6043	105	5938	56.55
G (GO annotation)	5837	462	5375	11.63	6043	518	5525	10.67
T (tumor)	5837	36	5801	161.14	6043	20	6023	301.15

Table 10 - Values of u_r for subtasks.

Subtask	**u_r**
A (alelle)	17
E (expression)	64
G (GO annotation)	11
T (tumor)	231

A common question that emerged was, what resources can be legitimately used to aid in categorizing the documents? In general, groups could use anything, including resources on the MGI Web site. The only resource they could not use was the direct data itself, i.e., data that was directly linked to the PMID or the associated MGI unique identifier. Thus, they could not go into the MGI database (or any other aggregated resource such as Entrez Gene or SOURCE) and pull out GO codes, tumor terms, mutant phenotypes, or any other data that was explicitly linked to a document. But anything else was fair game.

3.5 Results

A total of 46-48 runs were submitted for each of the four tasks. The results varied widely by subtask. The highest results were obtained in the tumor subtask, followed by the allele and expression subtasks very close to each other, and the GO subtask substantially lower. In light of the concern about the GO subtask and the inability of any feature beyond the MeSH term *Mice* to improve performance in 2004, this year's results are reassuring that document triage can potentially be helpful to model organism database curators. Table 11 shows the best and median U_{norm} values. Tables 12-15 show the results of the four subtasks; Figures 6-9 depict these results graphically.

From these results, it is clear that the GO task is somewhat different than the other tasks. The best utility scores that participants were able to achieve were in the 0.50-0.60 range, which were much lower than for the other three tasks. Another interesting observation is the u_r factor for the best performing task, tumor biology at 231, was the highest among the tasks, while the lowest occurred for the worst performing task, GO, at 11. While a high u_r leads to an increasing preference for high recall over precision, a u_r of 11 is still substantial compared to typical, more balanced classification tasks where the goal is often to optimize F-measure. Further investigation is needed to understand why the GO task appears more difficult than the other three. A separate analysis of the similar 2004 data shows that the individual GO codes are very sparsely represented in the training and test collections. This observation combined with assuming that correctly categorizing a paper is highly dependent upon the specific GO codes associated with the paper may explain why the GO task is more heterogeneous and therefore complex than the other tasks [6].

Table 11 - Best and median results for each subtask.

Subtask	**Best u_{norm}**	**Median u_{norm}**	**u_r**
A (alelle)	0.871	0.7773	17
E (expression)	0.8711	0.6413	64
G (GO annotation)	0.587	0.4575	11
T (tumor)	0.9433	0.761	231

4. Future Directions

The TREC Genomics 2005 Genomics Track was again carried out with much participation and enthusiasm. To prepare for the 2006 track, we created an on-line survey for members of the track email list. A total of 26 people responded to the survey, the results of which can be found at http://ir.ohsu.edu/genomics/2005survey.html. In summary, the results indicate that there is a strong desire for full-text journal articles for the ad hoc task and an information extraction task as the second task for the track in 2006.

Acknowledgements

The TREC Genomics Track is funded by grant ITR-0325160 from the U.S. National Science Foundation. The following individuals carried out interviews with biologists to obtain topics for the ad hoc task: Marti Hearst, Laura Ross, Leonie IJzereef, Dina Demner, Sharon Yang, Phoebe Roberts, William Cohen, Kevin Cohen, Paul Thompson, LV Subramaniam, and Jay Urbain. The track also thanks Ellen Voorhees, Ian Soboroff, and Lori Buckland of NIST for their help in various ways.

Table 12 - Results of allele subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
aibmadz05s [10]	ibm.zhang	0.4669	0.9337	0.6225	0.871
ABBR003SThr [42]	ibm.kanungo	0.4062	0.9458	0.5683	0.8645
ABBR003 [42]	ibm.kanungo	0.3686	0.9548	0.5319	0.8586
aibmadz05m1 [10]	ibm.zhang	0.5076	0.9006	0.6493	0.8492
aibmadz05m2 [10]	ibm.zhang	0.5025	0.9006	0.6451	0.8482
cuhkrun3A [43]	cuhk.lam	0.3442	0.9548	0.506	0.8478
THUIRgenA1p1 [19]	tsinghua.ma	0.4902	0.9006	0.6348	0.8455
cuhkrun2A [43]	cuhk.lam	0.3316	0.9578	0.4926	0.8443
aFduMarsII [29]	fudan.niu	0.4195	0.9187	0.576	0.8439
aNTUMAC [24]	ntu.chen	0.3439	0.9488	0.5048	0.8423
aFduMarsI [29]	fudan.niu	0.4754	0.9006	0.6223	0.8421
ASVMN03 [42]	ibm.kanungo	0.4019	0.9127	0.558	0.8327
aNLMB [12]	nlm-umd.aronson	0.3391	0.9398	0.4984	0.832
aDIMACSl9w [44]	rutgers.dayanik	0.4357	0.8976	0.5866	0.8292
THUIRgA0p9x [19]	tsinghua.ma	0.5414	0.8675	0.6667	0.8242
cuhkrun1 [43]	cuhk.lam	0.3257	0.9367	0.4833	0.8226
aDIMACSg9md [44]	rutgers.dayanik	0.4509	0.8855	0.5976	0.8221
aDIMACSl9md [44]	rutgers.dayanik	0.3844	0.9066	0.5399	0.8212
aDIMACSg9w [44]	rutgers.dayanik	0.4882	0.8705	0.6255	0.8168
NLM2A [12]	nlm-umd.aronson	0.4332	0.8795	0.5805	0.8118
AOHSUVP [23]	ohsu.hersh	0.3556	0.8976	0.5094	0.8019
aFduMarsIII [29]	fudan.niu	0.3254	0.9096	0.4794	0.7987
aDUTCat1 [18]	dalianu.yang	0.2858	0.9307	0.4374	0.7939
AOHSUSL [23]	ohsu.hersh	0.3448	0.8765	0.4949	0.7785
aQUT14 [45]	queensu.shatkay	0.3582	0.8675	0.507	0.776
AOHSUBF [23]	ohsu.hersh	0.3007	0.8976	0.4505	0.7748
aIBMIRLrul [46]	ibm-india.ramakrishnan	0.3185	0.8855	0.4685	0.7741
Ameta [47]	uwisconsin.craven	0.3031	0.8946	0.4527	0.7736
Apars [47]	uwisconsin.craven	0.2601	0.9277	0.4063	0.7725
aIBMIRLsvm [46]	ibm-india.ramakrishnan	0.2982	0.8946	0.4473	0.7707
aDUTCat2 [18]	dalianu.yang	0.262	0.9217	0.408	0.769
aMUSCUIUC3 [11]	uiuc.zhai	0.4281	0.8072	0.5595	0.7438
Afull [47]	uwisconsin.craven	0.2718	0.8825	0.4156	0.7434
aMUSCUIUC2 [11]	uiuc.zhai	0.5501	0.7771	0.6442	0.7397
aQUNB8 [45]	queensu.shatkay	0.3182	0.8464	0.4626	0.7397
aIBMIRLmet [46]	ibm-india.ramakrishnan	0.32	0.8434	0.464	0.738
ABPLUS [20]	erasmus.kors	0.241	0.8916	0.3795	0.7264
aUCHSCnb1En3 [35]	ucolorado.cohen	0.508	0.7651	0.6106	0.7215
aQUT11 [45]	queensu.shatkay	0.3785	0.7741	0.5084	0.6993
aUCHSCnb1En4 [35]	ucolorado.cohen	0.6091	0.6476	0.6277	0.6231
aMUSCUIUC1 [11]	uiuc.zhai	0.6678	0.6054	0.6351	0.5877
aUCHSCsvm [35]	ucolorado.cohen	0.7957	0.4458	0.5714	0.4391
aNLMF [12]	nlm-umd.aronson	0.2219	0.5301	0.3129	0.4208
LPC6	langpower.yang	0.4281	0.4307	0.4294	0.3969
FTA [20]	erasmus.kors	0.3562	0.3916	0.373	0.3499
aLRIk1	uparis-sud.kodratoff	0.2331	0.259	0.2454	0.2089
aLRIk3	uparis-sud.kodratoff	0.2191	0.262	0.2387	0.2071
aLRIk2	uparis-sud.kodratoff	0.2306	0.25	0.2399	0.2009
Minimum		0.2191	0.25	0.2387	0.2009
Median		0.3572	0.8931	0.5065	0.77725
Maximum		0.7957	0.9578	0.6667	0.871

Table 13 - Results of expression subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
eFduMarsI [29]	fudan.niu	0.1899	0.9333	0.3156	0.8711
eFduMarsII [29]	fudan.niu	0.1899	0.9333	0.3156	0.8711
eDUTCat1 [18]	dalianu.yang	0.1364	0.9429	0.2383	0.8496
eDIMACSl9w [44]	rutgers.dayanik	0.2026	0.9048	0.331	0.8491
eibmadz05s [10]	ibm.zhang	0.1437	0.9333	0.249	0.8464
eibmadz05m2 [10]	ibm.zhang	0.2109	0.8857	0.3407	0.8339
cuhkrun2E [43]	cuhk.lam	0.126	0.9333	0.222	0.8321
cuhkrun3E [43]	cuhk.lam	0.1481	0.9143	0.255	0.8321
EBBR0006SThr [42]	ibm.kanungo	0.1228	0.9333	0.2171	0.8292
THUIRgenE1p8 [19]	tsinghua.ma	0.1322	0.9238	0.2312	0.829
eibmadz05m1 [10]	ibm.zhang	0.2201	0.8762	0.3518	0.8277
EBBR0006 [42]	ibm.kanungo	0.1211	0.9333	0.2144	0.8275
eDUTCat2 [18]	dalianu.yang	0.1104	0.9429	0.1976	0.8241
ESVMN075 [42]	ibm.kanungo	0.1265	0.9143	0.2222	0.8156
cuhkrun1E [43]	cuhk.lam	0.1119	0.9143	0.1994	0.8009
eDIMACSg9w [44]	rutgers.dayanik	0.2444	0.8381	0.3785	0.7976
eFduMarsIII [29]	fudan.niu	0.0794	0.9524	0.1466	0.7799
eNTUMAC [24]	ntu.chen	0.1593	0.819	0.2667	0.7515
Epars [47]	uwisconsin.craven	0.0818	0.8857	0.1498	0.7304
eIBMIRLsvm [46]	ibm-india.ramakrishnan	0.0571	0.9238	0.1075	0.6854
ABPLUSE [20]	erasmus.kors	0.0841	0.819	0.1525	0.6796
eDIMACSg9md [44]	rutgers.dayanik	0.1575	0.7333	0.2593	0.672
Emeta [47]	uwisconsin.craven	0.1273	0.7333	0.2169	0.6548
eDIMACSl9md [44]	rutgers.dayanik	0.1054	0.7238	0.184	0.6278
NLM2E [12]	nlm-umd.aronson	0.2863	0.6381	0.3953	0.6132
EOHSUBF [23]	ohsu.hersh	0.0405	0.9619	0.0777	0.6058
Efull [47]	uwisconsin.craven	0.0636	0.781	0.1176	0.6012
EOHSUVP [23]	ohsu.hersh	0.0693	0.7429	0.1267	0.5869
EOHSUSL [23]	ohsu.hersh	0.0365	0.9905	0.0705	0.5824
eIBMIRLmet [46]	ibm-india.ramakrishnan	0.0627	0.7333	0.1155	0.5621
eIBMIRLrul [46]	ibm-india.ramakrishnan	0.0642	0.7238	0.1179	0.5589
eQUNB11 [45]	queensu.shatkay	0.1086	0.6381	0.1856	0.5563
eQUT18 [45]	queensu.shatkay	0.0967	0.5238	0.1632	0.4473
eMUSCUIUC1 [11]	uiuc.zhai	0.2269	0.4667	0.3053	0.4418
eMUSCUIUC3 [11]	uiuc.zhai	0.1572	0.4762	0.2364	0.4363
eQUNB19 [45]	queensu.shatkay	0.1132	0.4571	0.1815	0.4012
eUCHSCnb1En4 [35]	ucolorado.cohen	0.52	0.3714	0.4333	0.3661
FTE [20]	erasmus.kors	0.0835	0.4095	0.1387	0.3393
eUCHSCnb1En3 [35]	ucolorado.cohen	0.5714	0.3429	0.4286	0.3388
eNLMF [12]	nlm-umd.aronson	0.129	0.2286	0.1649	0.2045
eNLMKNN [12]	nlm-umd.aronson	0.0519	0.2381	0.0852	0.1701
eLRIk3	uparis-sud.kodratoff	0.0828	0.1238	0.0992	0.1024
eLRIk1	uparis-sud.kodratoff	0.1026	0.1143	0.1081	0.0987
eLRIk2	uparis-sud.kodratoff	0.1026	0.1143	0.1081	0.0987
eUCHSCsvm [35]	ucolorado.cohen	1	0.0381	0.0734	0.0381
eMUSCUIUC2 [11]	uiuc.zhai	0	0	0	-0.0074
Minimum		0	0	0	-0.0074
Median		0.12195	0.8	0.1985	0.6413
Maximum		1	0.9905	0.4333	0.8711

Table 14 - Results of GO subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
gFduMarsII [29]	fudan.niu	0.2122	0.8861	0.3424	0.587
gFduMarsI [29]	fudan.niu	0.2644	0.778	0.3947	0.5813
gIBMIRLmet [46]	ibm-india.ramakrishnan	0.2028	0.9015	0.3311	0.5793
gDUTCat1 [18]	dalianu.yang	0.1914	0.9286	0.3174	0.572
gIBMIRLrul [46]	ibm-india.ramakrishnan	0.1883	0.9286	0.3132	0.5648
gIBMIRLsvm [46]	ibm-india.ramakrishnan	0.2069	0.8668	0.3341	0.5648
gFduMarsIII [29]	fudan.niu	0.191	0.9093	0.3157	0.5591
GBBR004 [42]	ibm.kanungo	0.1947	0.8938	0.3198	0.5577
GOHSUBF [23]	ohsu.hersh	0.1889	0.9093	0.3127	0.5542
GAbsBBR0083 [42]	ibm.kanungo	0.2524	0.7548	0.3783	0.5516
GOHSUVP [23]	ohsu.hersh	0.2308	0.7819	0.3564	0.5449
GSVMN08 [42]	ibm.kanungo	0.2038	0.8436	0.3283	0.5441
gDUTCat2 [18]	dalianu.yang	0.1779	0.9363	0.2989	0.5428
gNTUMAC [24]	ntu.chen	0.1873	0.8803	0.3089	0.5332
gibmadz05m2 [10]	ibm.zhang	0.3179	0.6216	0.4206	0.5004
gibmadz05m1 [10]	ibm.zhang	0.3216	0.6178	0.423	0.4993
ABPLUSG [20]	erasmus.kors	0.2178	0.7259	0.3351	0.4889
gDIMACSl9w [44]	rutgers.dayanik	0.245	0.668	0.3585	0.4809
gibmadz05s [10]	ibm.zhang	0.3226	0.583	0.4154	0.4717
GOHSUSL [23]	ohsu.hersh	0.2536	0.6429	0.3637	0.4709
gDIMACSl9md [44]	rutgers.dayanik	0.2425	0.6564	0.3542	0.47
cuhkrun1G [43]	cuhk.lam	0.2706	0.6139	0.3757	0.4635
gDIMACSg9md [44]	rutgers.dayanik	0.2529	0.6293	0.3608	0.4603
NLM2G [12]	nlm-umd.aronson	0.3223	0.5656	0.4107	0.4575
Gpars [47]	uwisconsin.craven	0.1862	0.7587	0.299	0.4572
gDIMACSg9w [44]	rutgers.dayanik	0.2754	0.5965	0.3768	0.4538
THUIRgenGMNG [19]	tsinghua.ma	0.2107	0.6776	0.3214	0.4468
Gmeta [47]	uwisconsin.craven	0.1689	0.7934	0.2785	0.4386
NLM1G [12]	nlm-umd.aronson	0.316	0.5405	0.3989	0.4342
cuhkrun2G [43]	cuhk.lam	0.2109	0.6506	0.3185	0.4293
Gfull [47]	uwisconsin.craven	0.1904	0.6988	0.2993	0.4287
THUIRgenG1p1 [19]	tsinghua.ma	0.1827	0.6506	0.2852	0.3859
gQUNB15 [45]	queensu.shatkay	0.2102	0.5676	0.3067	0.3736
gNLMF [12]	nlm-umd.aronson	0.1887	0.6062	0.2878	0.3693
gQUT22 [45]	queensu.shatkay	0.1811	0.6158	0.2799	0.3628
gQUNB12 [45]	queensu.shatkay	0.1603	0.6602	0.258	0.3459
cuhkrun3G [43]	cuhk.lam	0.1651	0.5637	0.2554	0.3045
gUCHSCnb1En3 [35]	ucolorado.cohen	0.4234	0.3417	0.3782	0.2994
gMUSCUIUC1 [11]	uiuc.zhai	0.393	0.2799	0.3269	0.2406
FTG [20]	erasmus.kors	0.2211	0.2876	0.25	0.1955
gUCHSCnb1En4 [35]	ucolorado.cohen	0.5542	0.1776	0.269	0.1646
gUCHSCsvm [35]	ucolorado.cohen	0.406	0.1834	0.2527	0.159
gMUSCUIUC3 [11]	uiuc.zhai	0.0891	0.3456	0.1416	0.0242
gLRIk3	uparis-sud.kodratoff	0.0998	0.1158	0.1072	0.0209
gLRIk2	uparis-sud.kodratoff	0.1	0.1023	0.1011	0.0186
gLRIk1	uparis-sud.kodratoff	0.0938	0.1023	0.0979	0.0125
gMUSCUIUC2 [11]	uiuc.zhai	0.0706	0.1737	0.1004	-0.0342
Minimum		0.0706	0.1023	0.0979	-0.0342
Median		0.2102	0.6506	0.3185	0.4575
Maximum		0.5542	0.9363	0.423	0.587

Table 15 - Results of tumor subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
tDIMACSg9w [44]	rutgers.dayanik	0.0709	1	0.1325	0.9433
TSVM0035 [42]	ibm.kanungo	0.0685	1	0.1282	0.9411
tDIMACSg9md [44]	rutgers.dayanik	0.0556	1	0.1053	0.9264
tFduMarsI [29]	fudan.niu	0.1061	0.95	0.191	0.9154
tFduMarsII [29]	fudan.niu	0.099	0.95	0.1792	0.9126
tIBMIRLmet [46]	ibm-india.ramakrishnan	0.0945	0.95	0.1719	0.9106
tDIMACSl9w [44]	rutgers.dayanik	0.0444	1	0.0851	0.9069
TBBR0004SThr [42]	ibm.kanungo	0.0436	1	0.0835	0.905
cuhkrun3T [43]	cuhk.lam	0.0426	1	0.0818	0.9028
tibmadz05m2 [10]	ibm.zhang	0.0757	0.95	0.1402	0.8998
tDUTCat1 [18]	dalianu.yang	0.0745	0.95	0.1382	0.8989
tibmadz05s [10]	ibm.zhang	0.0688	0.95	0.1284	0.8944
tibmadz05m1 [10]	ibm.zhang	0.0674	0.95	0.1258	0.8931
TBBR0004 [42]	ibm.kanungo	0.0376	1	0.0725	0.8892
tDUTCat2 [18]	dalianu.yang	0.035	1	0.0677	0.8807
tNTUMACwj [24]	ntu.chen	0.0518	0.95	0.0982	0.8747
tIBMIRLrul [46]	ibm-india.ramakrishnan	0.0415	0.95	0.0795	0.855
cuhkrun1T [43]	cuhk.lam	0.0769	0.9	0.1417	0.8532
tFduMarsIII [29]	fudan.niu	0.0286	1	0.0556	0.8528
tNTUMAC [24]	ntu.chen	0.0526	0.9	0.0994	0.8299
tDIMACSl9md [44]	rutgers.dayanik	0.0323	0.95	0.0625	0.8268
Tpars [47]	uwisconsin.craven	0.0317	0.95	0.0613	0.8242
ABPLUST [20]	erasmus.kors	0.0314	0.95	0.0607	0.8229
Tfull [47]	uwisconsin.craven	0.0443	0.9	0.0845	0.816
Tmeta [47]	uwisconsin.craven	0.0523	0.85	0.0986	0.7833
THUIRgenT1p5 [19]	tsinghua.ma	0.0213	0.95	0.0417	0.761
TOHSUSL [23]	ohsu.hersh	0.0254	0.9	0.0493	0.7502
tQUNB3 [45]	queensu.shatkay	0.0244	0.9	0.0474	0.7439
TOHSUBF [23]	ohsu.hersh	0.0192	0.95	0.0376	0.7396
TOHSUVP [23]	ohsu.hersh	0.0237	0.9	0.0462	0.7394
tMUSCUIUC3 [11]	uiuc.zhai	0.3182	0.7	0.4375	0.6935
tIBMIRLsvm [46]	ibm-india.ramakrishnan	0.0308	0.8	0.0593	0.6909
tQUT10 [45]	queensu.shatkay	0.0132	1	0.026	0.6758
tMUSCUIUC2 [11]	uiuc.zhai	0.0828	0.7	0.1481	0.6665
tQUT14 [45]	queensu.shatkay	0.3095	0.65	0.4194	0.6437
NLM1T [12]	nlm-umd.aronson	0.0813	0.65	0.1444	0.6182
NLM2T [12]	nlm-umd.aronson	0.0813	0.65	0.1444	0.6182
tMUSCUIUC1 [11]	uiuc.zhai	0.3429	0.6	0.4364	0.595
tNTUMACasem [24]	ntu.chen	0.0339	0.65	0.0645	0.5699
LPC7	langpower.yang	0.3548	0.55	0.4314	0.5457
FTT [20]	erasmus.kors	0.0893	0.5	0.1515	0.4779
tNLMF [12]	nlm-umd.aronson	0.0207	0.55	0.0399	0.4372
cuhkrun2T [43]	cuhk.lam	0.0268	0.4	0.0503	0.3372
tUCHSCnb1En3 [35]	ucolorado.cohen	0.1935	0.3	0.2353	0.2946
tUCHSCnb1En4 [35]	ucolorado.cohen	0.375	0.15	0.2143	0.1489
tLRIk2	uparis-sud.kodratoff	0.0909	0.1	0.0952	0.0957
tLRIk1	uparis-sud.kodratoff	0.087	0.1	0.093	0.0955
tLRIk3	uparis-sud.kodratoff	0.069	0.1	0.0816	0.0942
tUCHSCsvm [35]	ucolorado.cohen	1	0.05	0.0952	0.05
Tcsusm2 [30]	csusm.guillen	0.0256	0.05	0.0339	0.0418
Tcsusm1 [30]	csusm.guillen	0.0244	0.05	0.0328	0.0413
Minimum		0.0132	0.05	0.026	0.0413
Median		0.0526	0.9	0.0952	0.761
Max		1	1	0.4375	0.9433

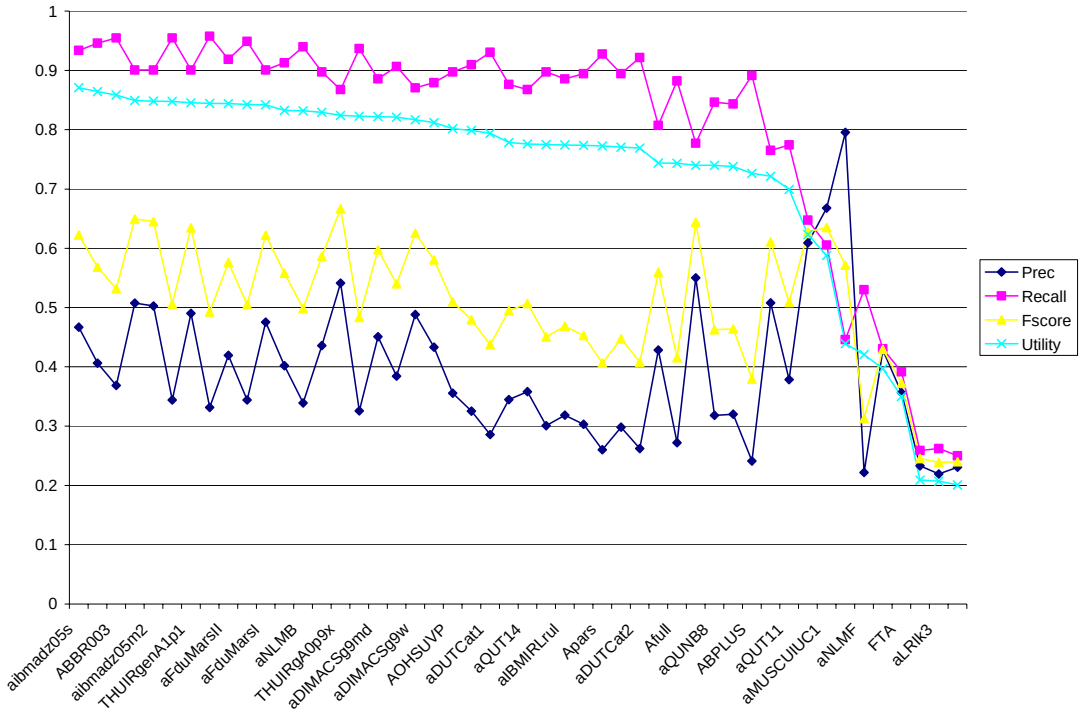

Figure 6 - Results of allele subtask by run displayed graphically, sorted by utility measure.

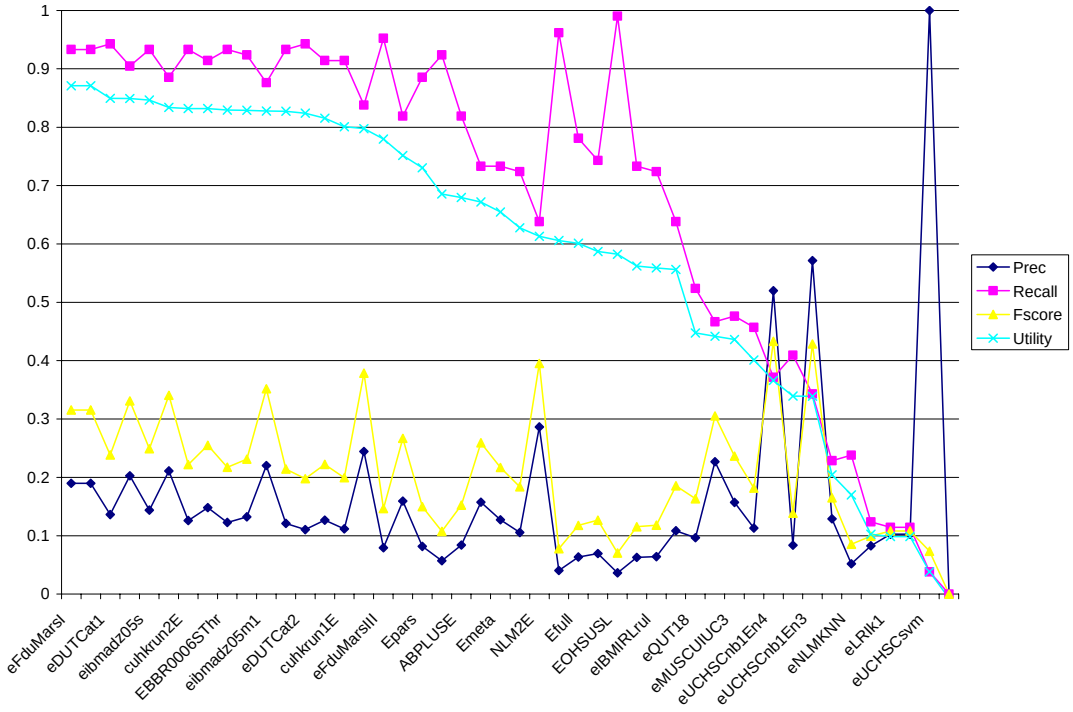

Figure 7 - Results of expression subtask by run displayed graphically, sorted by utility measure.

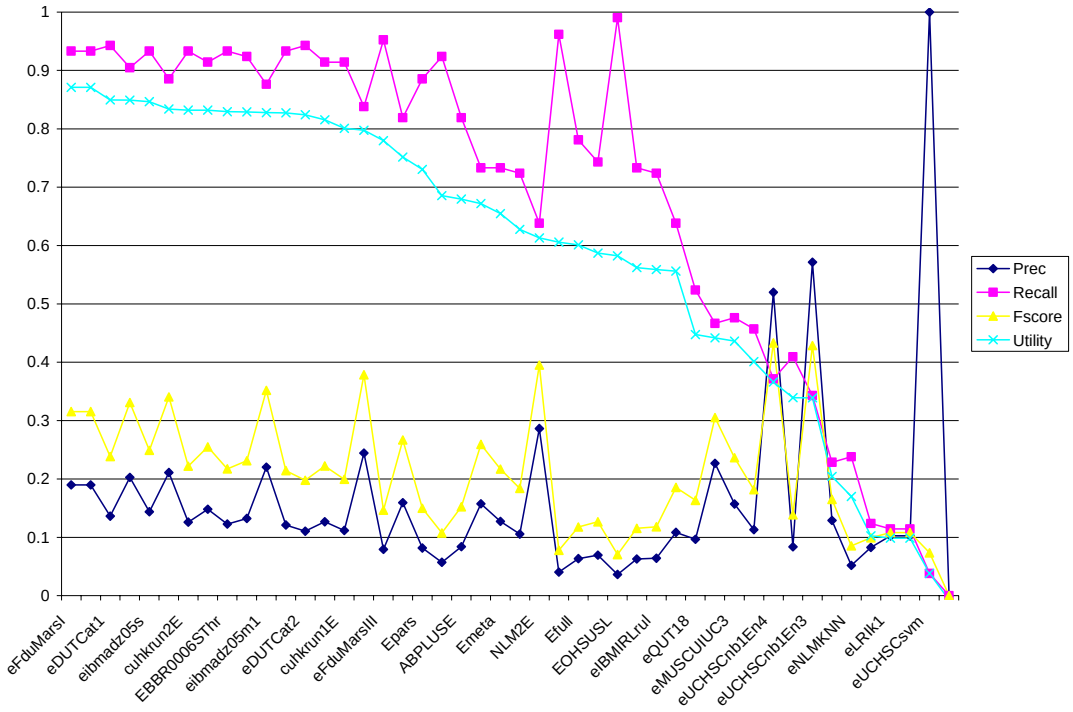

Figure 8 - Results of GO subtask by run displayed graphically, sorted by utility measure.

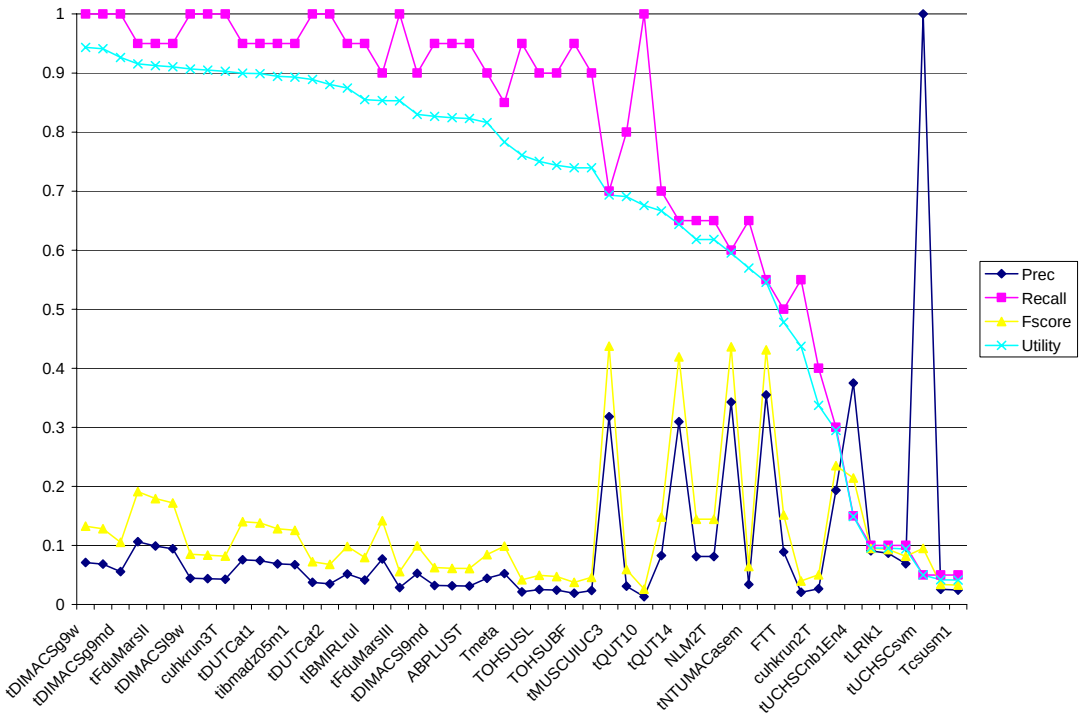

Figure 9 - Results of tumor subtask by run displayed graphically, sorted by utility measure.

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. Cohen AM and Hersh WR, A survey of current work in biomedical text mining. Briefings in Bioinformatics, 2005. 6: 57-71.
4. Hersh WR and Bhupatiraju RT. TREC genomics track overview. The Twelfth Text Retrieval Conference (TREC 2003). 2003. Gaithersburg, MD: NIST. 14-23. http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.
5. Hersh WR, et al., Enhancing access to the bibliome: the TREC 2004 Genomics Track. Journal of Biomedical Discovery and Collaboration, 2006: in press.
6. Cohen AM and Hersh WR, The TREC 2004 Genomics Track categorization task: classifying full-text biomedical documents. Journal of Biomedical Discovery and Collaboration, 2006: in press.
7. Cohen AM, et al., Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association, 2006: in press.
8. Buckley C and Voorhees EM. Retrieval evaluation with incomplete information. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004. Sheffield, England: ACM Press. 25-32.
9. Huang X, Zhong M, and Si L. York University at TREC 2005: Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/yorku-huang2.geo.pdf.
10. Ando RK, Dredze M, and Zhang T. TREC 2005 Genomics Track experiments at IBM Watson. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ibm-tjwatson.geo.pdf.
11. Zhai C, et al. UIUC/MUSC at TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uillinois-uc.geo.pdf.
12. Aronson AR, et al. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/nlm-umd.geo.pdf.
13. Tsai TH, et al. Enhance genomic IR with term variation and expansion: experience of the IASL Group at Genomic Track 2005. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/academia-sinica.geo.pdf.
14. Abdou S, Savoy J, and Ruch P. Evaluation of stemming, query expansion and manual indexing approaches for the genomic task. The Fourteenth Text REtrieval Conference

Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uneuchatel.geo.pdf.
15. Ruch P, et al. Report on the TREC 2005 experiment: Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uhospital-geneva.geo.pdf.
16. Yu N, et al. TREC 2005 Genomics Track at I2R. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/inst-infocomm.geo.pdf.
17. Pirkola A. TREC 2005 Genomics Track experiments at UTA. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/utampere.geo.pdf.
18. Yang Z, et al. TREC 2005 Genomics Track experiments at DUTAI. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/dalianu.geo.pdf.
19. Li J, et al. Learning domain-specific knowledge from context - THUIR at TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/tsinghuau-ma.geo.pdf.
20. Schijvenaars BJA, et al. TREC 2005 Genomics Track - a concept-based approach to text categorization. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/erasmus-tno.geo.pdf.
21. Meij E, et al. Combining thesauri-based methods for biomedical retrieval. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uamsterdam-infoinst.geo.pdf.
22. Ruiz ME and Southwick SB. UB at CLEF 2005: bilingual CLIR and medical image retrieval tasks. 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Springer Lecture Notes in Computer Science. 2005. Vienna, Austria: Springer-Verlag. in press.

23. Cohen AM, Yang J, and Hersh WR. A comparison of techniques for classification and ad hoc retrieval of biomedical documents. The Fourteenth Text Retrieval Conference - TREC 2005. 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ohsu-geo.pdf.
24. Lin KHY, Hou WJ, and Chen HH. Retrieval of biomedical documents by prioritizing key phrases. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ntu.geo.adhoc.pdf.
25. Bacchin M and Melucci M. Symbol-based query expansion experiments at TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/upadova.geo.pdf.

26. Urbain J, Goharian N, and Frieder O. IIT TREC 2005: Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/iit-urbain.geo.pdf.
27. Camous F, et al. Structural term extraction for expansion of template-based genomic queries. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/dublincu.geo.pdf.
28. Shi Z, et al. Synonym-based expansion and boosting-based re-ranking: a two-phase approach for genomic information retrieval. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/simon-fraseru.geo.pdf.
29. Niu J, et al. WIM at TREC 2005. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/fudanu-sun.geo.ent.pdf.
30. Guillen R. CSUSM at TREC 2005: Genomics and Enterprise Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/calstateu-sanmarcos.geo.ent.pdf.
31. Zhou W and Yu C. Experiment report of TREC 2005 Genomics Track ad hoc retrieval task. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uillinois-chicago.geo.pdf.
32. Lin J, et al. A menagerie of tracks at Maryland: HARD, Enterprise, QA, and Genomics, oh my! The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/umaryland-lin.hard.ent.qa.geo.pdf.
33. Eichmann D and Srinivasan P. Experiments in questions and relationships at the University of Iowa. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uiowa.geo.qa.pdf.
34. Huang L, Chen Z, and Murphey YL. UM-D at TREC 2005: Genomics Track. The Fourteenth Text Retrieval Conference - TREC 2005. 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/umich-dearborn.geo.pdf.
35. Caporaso JG, et al. Concept recognition and the TREC genomics tasks. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ucolorado-hsc.geo.pdf.
36. Zakharov M. DataparkSearch at TREC 2005. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/datapark.geo.pdf.
37. Eppig JT, et al., The Mouse Genome Database (MGD): from genes to mice - a community resource for mouse biology. Nucleic Acids Research, 2005. 33: D471-D475.
38. Anonymous, The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 2004. 32: D258-D261.

39. Krupke D, et al., The Mouse Tumor Biology Database: integrated access to mouse cancer biology data. Experimental Lung Research, 2005. 31: 259-270.
40. Hill DP, et al., The mouse Gene Expression Database (GXD): updates and enhancements. Nucleic Acids Research, 2004. 32: D568-D571.
41. Strivens M and Eppig JT, Visualizing the laboratory mouse: capturing phenotype information. Genetica, 2004. 122: 89-97.
42. Si L and Kanungo T. Thresholding strategies for text classifiers: TREC 2005 biomedical triage task experiments. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/carnegie-mu-kanungo.geo.pdf.
43. Lam W, Han Y, and Chan K. Pattern-based customized learning for TREC Genomics Track categorization task. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/chineseu-hongkong-lam.geo.pdf.
44. Dayanik A, et al. DIMACS at the TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/rutgersu-dimacs.geo.pdf.
45. Zheng ZH, et al. Applying probabilistic thematic clustering for classification in the TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/queensu.geo.pdf.
46. Subramaniam LV, Mukherjea S, and Punjani D. Biomedical document triage: automatic classification exploiting category specific knowledge. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ibm-india.subramaniam.geo.pdf.
47. Brow T, Settles B, and Craven M. Classifying biomedical articles by making localized decisions. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uwisconsin.geo.pdf.