TREC 2005 Genomics Track

William Hersh
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu

These slides and track information at
http://ir.ohsu.edu/genomics

TREC Genomics Track plenary session

10:00-10:30	Overview – Bill Hersh – Oregon Health & Science University
10:30-11:00	BREAK
11:00-11:20	National Library of Medicine
11:20-11:40	IBM Ando
11:40-12:00	York University
12:00-12:20	Rutgers University DIMACS

Don't forget track workshop tomorrow at noon.

Acknowledgements

- Track participants and volunteers
  - Topic collectors acknowledged in paper
- OHSU team
  - Data management – Ravi Teja Bhupatiraju
  - Analysis – Aaron Cohen, Jianji Yang
  - Relevance judges – Laura Ross, Phoebe Roberts, Andrew Amata, Alita Miller, Bradley Feilmeier
- Data providers
  - National Library of Medicine
  - Mouse Genomic Informatics
- Funder
  - National Science Foundation Grant ITR-0325160
- NIST and Ellen Voorhees
- Track steering committee

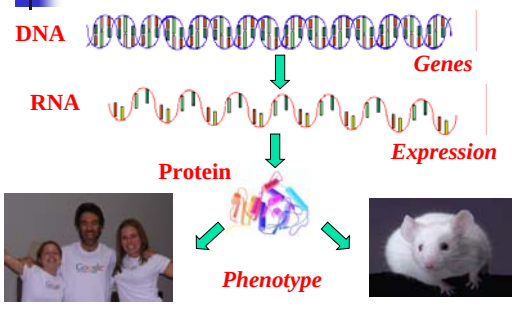
Overview of talk

- Motivations
- Past Results
- TREC Genomics 2005 Track
  - Tasks
  - Measures
  - Results
- Future Directions

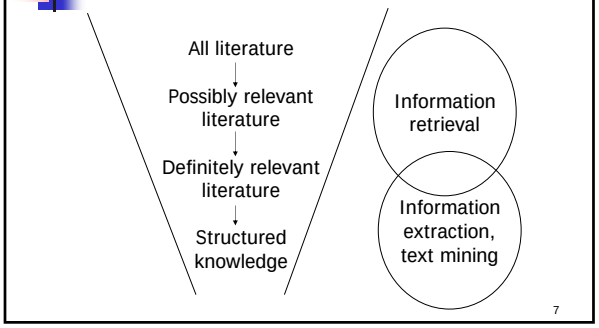
Motivation

- We are in an era of "high throughput", data-intensive science
- Biology and medicine provide many information challenges for information retrieval, extraction, mining, etc.
- Many reasons to structure knowledge with development of annotation, model organism databases, cross-data linkages, etc.
- Growing array of publicly accessible data resources and tools that may aid these tasks

Basic biology primer – but it's really not quite this simple



Emerging approach to biological knowledge management



7

TREC 2003 Genomics Track

- Constrained by lack of resources
  - Aided by Gene Reference into Function (GeneRIF) annotations in LocusLink (now Entrez Gene), which are linked to PubMed IDs
- Ad hoc document retrieval task
  - Searching MEDLINE documents for articles about function of a gene, with GeneRIFs as relevance judgments
- Extraction task
  - Identifying text of GeneRIF
  - Assessed by string overlap – Dice and derivatives

8

TREC 2004 Genomics Track

- Ad hoc retrieval task
  - Modeled after biologist with acute information needs
  - Used MEDLINE bibliographic database – despite proliferation of full-text journals, still entry point into literature for most searchers
- Categorization tasks
  - Motivated by real-world problems faced by Mouse Genome Informatics (MGI) curators, e.g., choosing articles and applying Gene Ontology (GO) terms for gene function
  - Divided into subtasks of article triage and gene/article/GO hierarchy annotation

9

TREC 2005 Genomics Track

- An "incremental" year – used variants of 2004 track with same underlying document collections
- Two tasks
  - Ad hoc retrieval
  - Categorization

10

Participation – continues to grow; largest in TREC

Year	Groups doing ad hoc task	Groups doing "other" task	Total groups
2003	25	14	28
2004	27	20	33
2005	32	19	41

11

Ad hoc retrieval task

- Documents
- Topics
- Relevance judgments
- Results
- Preliminary analysis

12

Ad hoc retrieval task documents

- Continued use of MEDLINE subset
  - 10 years from 1994 to 2003
  - ~4.5M documents
    - About one-third of entire database, which goes back to 1966
  - ~9 GB text (MEDLINE format)
- Note: promoting use of collection for other tasks beyond Genomics Track

13

Ad hoc retrieval task topics

- Instead of general information needs statements, decided to focus on more structured topics
- Still representative of common information needs but might allow other resources to be used to improve results
- Developed "generic topic types" (GTTs) and then interviewed real biologists to obtain real information needs that fit into template
- Transformed information needs into searchable topics

14

GTTs

Generic Topic Type (GTT)	Range
Find articles describing standard methods or protocols for doing some sort of experiment or procedure	100-109
Find articles describing the role of a gene involved in a given disease	110-119
Find articles describing the role of a gene in a specific biological process	120-129
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
Find articles describing one or more mutations of a given gene and its biological impact	140-149

15

Example topics for selected GTTs

Generic Topic Type (GTT)	Example Topic
Find articles describing standard methods or protocols for doing some sort of experiment or procedure	Method or protocol: GST fusion protein expression in Sf9 insect cells
Find articles describing the role of a gene in a specific biological process	Gene: Insulin receptor gene Biological process: Signaling tumorigenesis
Find articles describing one or more mutations of a given gene and its biological impact	Gene with mutation: Ret Biological impact: Thyroid function

16

Relevance judgments

- Using usual TREC pooling method
  - Assessed top designated runs of the 27 groups who submitted results
- Performed by five judges with varying expertise in biology
- Averages per topic
  - Documents assessed: 820 (total 41,018)
  - Definitely relevant: 50.5 (6.2%; range 0-527)
  - Possibly relevant: 41.2 (5.0%; range 0-182)
  - Definitely + possibly relevant (relevance for runs): 91.7 (11.2%; range 0-709)
- One topic (135) had no definitely or possibly relevant documents, so omitted from analysis

17

Judgment consistency

Judge 2 Judge 1	Relevant	Not relevant	Total
Relevant	1100	629	1729
Not relevant	546	8204	8750
Total	1646	8833	10479

- 9 topics judged in duplicate
- One topic judged three times
- Kappa = 0.58 -> "fair" agreement

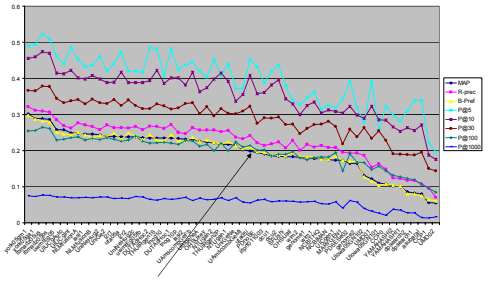
18

Metrics and analysis

- Primary performance metric – mean average precision (MAP)
- Also measured B-Pref, R-Prec, and precision@N documents
  - Original B-Pref results incorrect due to non-inclusion of nonrelevants in qrels
- Additional measurements provided by new version of trec_eval
- Statistical analysis – repeated measures ANOVA with posthoc pairwise comparisons
- Complete table of all official runs in paper

19

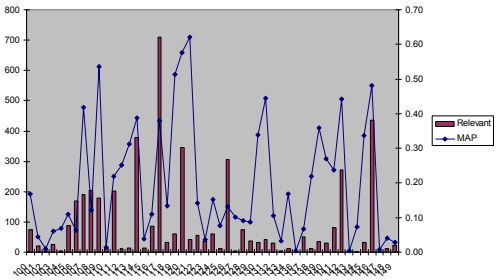
Ad hoc task results – sorted by MAP



Pairwise statistically significant from top run

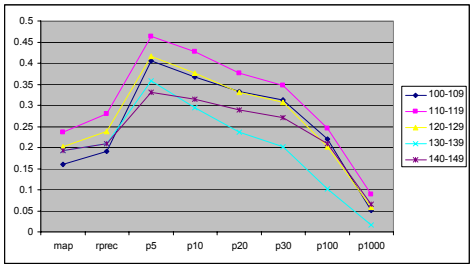
20

Ad hoc results – relevant and MAP by topic



21

Difference by GTT? Some, but not much…



22

Ad hoc task analysis – general observations so far

- Manual synonym expansion helps (York – best run with MAP of 0.3136), automated does not (IBM Watson, NLM)
- Relevance feedback without term expansion helps (UIUC)
- Basic Okapi with good parameters gives good baseline performance (several)
  - But better characterization of baseline experiments would improve our understanding

23

Categorization task

- Motivation
  - Apply text categorization to full-text documents for tasks that assist work of MGI
- Task
  - Decided to focus on document triage this year, keeping last year's one sub-task and adding three new ones
  - This type of task can have practical value in biomedicine

24

Triage subtasks – identifying articles for databases

- Alleles of mutant phenotypes – characteristics of organisms that have gene mutations
- Embryologic gene expression – which genes are expressed at various points in embryo development
- Gene Ontology – biological function of gene products
- Tumor biology – genes and mutations associated with development of tumors

25

Documents and categorization data

- Documents
  - Full text from three journals published by Highwire Press
  - Provided "crosswalk" to MEDLINE record
  - Created filtered subset for words "mouse", "mice", or "murine" – approach of MGI
- Partition of training and test data
  - 2002 – training data
  - 2003 – test data
- Triage status obtained from actual decisions by MGI
  - No internal (from track) relevance judgments
  - Participants could not use direct data but allowed to use other genomics information

26

Full-text documents for categorization task

Journal	2002 papers – total, subset	2003 papers – total, subset	Total papers – total, subset
J. of Biological Chemistry	6566, 4199	6593, 4282	13159, 8481
J. Of Cell Biology	530, 256	715, 359	1245, 615
Proceedings of NAS	3041, 1382	2888, 1402	5929, 2784
Total	10137, 5837	10196, 6043	20333, 11880

"Subset" papers – those with mouse, mice, or murine

27

Triage task performance was based on utility measurement

- $U_{norm} = U_{raw} / U_{max}$
- $U_{raw} = U_r$ * true positives + U_{nr} * false positives
  - U_r = relative utility of relevant document
  - U_{nr} = relative utility of nonrelevant document
- Set u_r based on boundary cases
  - $U_{nr} = -1$
  - $0.0 = U_r$ * all positives – all positives
  - U_r = all negatives / all positives
- Since varied across training and test data, set fixed value for each (rounded average of test & train)
- Utility clearly gives more weight to identifying relevants than omitting nonrelevants

28

Calculating u_r

Task	Training positive	Training negative	U_r calc	Test positive	Test negative	U_r calc	U_r actual
A	338	5499	16.3	332	5711	17.2	17
E	81	5756	71.1	105	5938	56.6	64
G	462	5375	11.6	518	5525	10.7	11
T	36	5801	161.4	20	6023	301.2	231

Documents: 5837 Documents: 6043

29

Annotation subtask measurements

Task	Best u_{norm}	Median u_{norm}	Observations
A	0.871	0.778	Middle range of performance and u_r
E	0.871	0.641	Middle range of performance and u_r
G	0.587	0.458	Little difference from last year; lowest u_r and still hardest
T	0.943	0.761	Highest u_r; fewest relevant

30

Triage task analysis

- Many different approaches used, with less-than-ideal reporting of baselines, so hard to compare
- Determination of proper threshold is essential in probability-based techniques
- Full text, MeSH terms, and *Mice* filtering helpful
- Binary feature weighting often as good or better than TF*IDF, cosine normalization

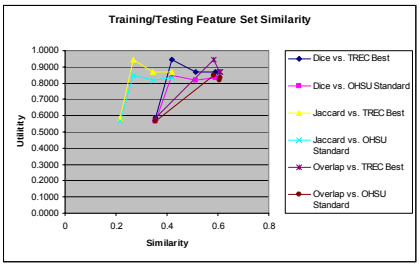
31

Training/Testing feature set – similarity vs. utility

- Computed similarity measures for feature sets computed by chi-square (alpha = 0.05) on training and test collections
  - Overlap (A, B) = |A U B| / min(|A|, |B|)
  - Dice (A, B) = 2*|A ∩ B| / (|A| + |B|)
  - Jaccard (A, B) = |A ∩ B| / |A U B|
- Lowest for GO, highest for Allele

32

Feature set similarity vs. utility



Trend towards correlation (overlap statistically significant)

33

Where do we go from here? Results of mailing list survey

- Posted August, 2005
- 26 respondents; results on Web site
- Clear preferences
  - Ad hoc retrieval should move to full text of journal articles
  - "Second" task should focus on information extraction, with some interest in question-answering and summarization

34

Future directions

- Continuation of track until (at least!) 2008, thanks to NSF grant
- Aim to develop enduring test collections from track data
  - Using MEDLINE collection for other tasks (Cohen, systematic drug efficacy reviews; Bernstam, important bibliographies)
- Future goals (from 2003 roadmap) include
  - Full-text retrieval
    - Obtaining full-text journals from various sources
    - Important to remember, however, that MEDLINE is still the entry point for most users to biomedical literature
  - Interactive user experiments
  - Broader types of users, information needs, tasks

35