TREC 2004 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://medir.ohsu.edu/~genomics

TREC Genomics Track plenary session

2:00-2:30	Overview; Bill Hersh – Oregon Health & Science University
2:30-2:50	Kazuhiro Seki – Indiana University
2:50-3:10	Sumio Fujita – Patolis Corp.
3:10-3:30	Aynur Dayanik – Rutgers University
3:30-3:50	Stefan Büttcher – University of Waterloo
4:10-5:30	Track Workshop

Acknowledgements

- Track participants and volunteers
- OHSU team
  - Data management – Ravi
  - Relevance judges – Laura, Phoebe
- Data providers
  - National Library of Medicine
  - Mouse Genomic Informatics
- Funder
  - National Science Foundation Grant ITR-0325160
- NIST and Ellen Voorhees
- Track steering committee

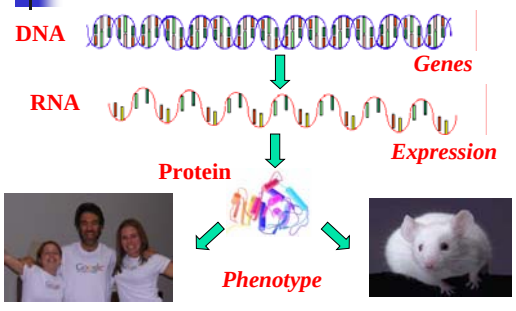
Overview of talk

- Motivations
- TREC Genomics 2004 Track
  - Tasks
  - Measures
  - Results
- Future Directions

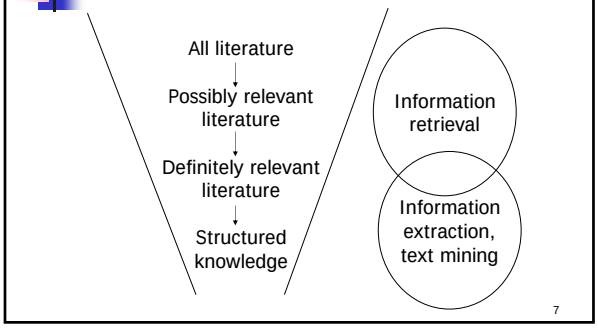
Motivation

- 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 but partially overcome by great enthusiasm
  - Aided by Gene Reference into Function (GeneRIF) annotations in LocusLink, which are linked to PubMed IDs
- Primary task – ad hoc document retrieval
  - Searching MEDLINE documents for articles about function of a gene, with GeneRIFs as relevance judgments
- Secondary task – identifying text of GeneRIF
  - Assessed by string overlap – Dice and derivatives
- Results and papers on TREC Web site

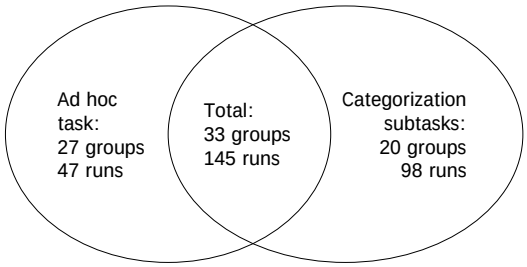
8

TREC 2004 Genomics Track

- Two tasks
  - Ad hoc retrieval
    - 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
    - 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 annotation

9

Track participation – largest number of participants



10

Ad hoc retrieval task

- Documents
  - 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)
- Topics
  - Based on real biologist information needs
  - 50 topics (and 5 samples) based on
    - 74 real information needs
    - Collected from 43 biologists by 11 interviewers
    - Each reviewed by 1-2 others who turned into "searchable" topic

11

Example topic

<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>

12

Relevance judgments

- Using usual TREC pooling method
  - Assessed top designated runs of the 27 groups who submitted results
- Performed by two judges – a PhD biologist and undergraduate biologist
  - Kappa = 0.51 – agreement "fair"; details in paper
- Averages per topic
  - Documents assessed: 975
  - Definitely relevant: 93 (9%; range 1-506)
  - Possibly relevant: 73 (7%; range 0-485)
  - Definitely + possibly relevant (relevance for runs): 166 (16%; range 1-697)
    - Three topics had no definitely relevant documents

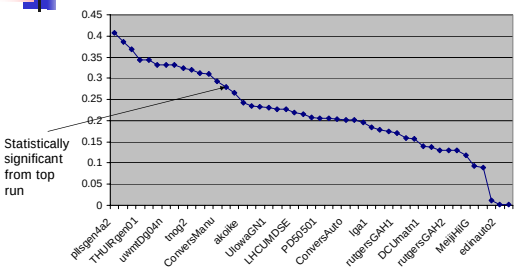
13

Metrics and analysis

- Primary performance metric – mean average precision (MAP)
- Also measured precision@10 and precision@100 documents
- Groups had additional measurements from trec_eval
- Statistical analysis – repeated measures ANOVA with posthoc Tukey pairwise comparisons
- Complete table of all official runs in paper

14

Ad hoc task results



n = 47, Max = .4075, Median = .2074, Min = .0012

15

Ad hoc task analysis

- Best runs used a variety of techniques, including
  - Domain-specific query expansion
  - Language modeling techniques, e.g., smoothing
- Of note, simple OHSU runs using Lucene "out of the box" (TF*IDF weighting) scored above mean/median
  - OHSUNeeds = .2343, OHSUAll = .2272
  - In other words, many groups did detrimental things!

16

Categorization task

- Motivation
  - Apply text categorization to full-text documents for tasks that assist work of MGI
- Sub-tasks
  - Triage – determine if articles have experimental evidence warranting GO assignment
    - A pertinent task beyond gene function annotation
  - Annotation – determine if article warrants assignment of GO category, with or without evidence code(s)
- Why not annotate actual GO terms?
  - Avoid exact overlap with Biocreative
  - A hard task, as learned from Biocreative

17

Gene Ontology (GO, www.geneontology.org)

- "Ontology" of ~18,000 terms reflecting gene function based on three hierarchies
  - Molecular function (MF)
  - Biological process (BP)
  - Cellular component (CC)
- Most model organism databases assign GO terms to genes linked to literature
- Evidence code denotes type of experimental support
  - Not all evidence is created equally, e.g., from nontraceable author statement (NAS) to inferred from direct assay (IDA)

18

Categorization task (cont.)

- 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
- Association with genes and GO codes
  - Data from MGI
  - No internal (from track) relevance judgments
- Partition of training and test data
  - 2002 – training data
  - 2003 – test data

19

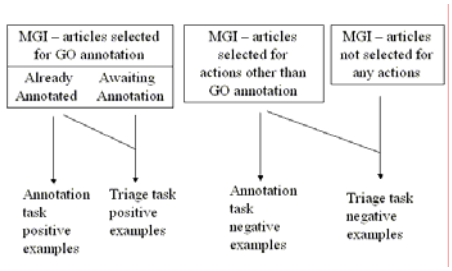
Full-text documents for categorization task

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

20

Partition of data for tasks



21

Triage subtask measurements

- Primary – normalized utility measure
  - $U_{norm} = U_{raw} / U_{max}$
  - U_{raw} = factor * true positives – false positives
  - Developed "desired" boundary cases
    - Completely perfect prediction: $U_{norm} = 1$
    - All documents positive (triage all): $1 > U_{norm} > 0$
    - All documents negative (triage none): $U_{norm} = 0$
    - Completely imperfect prediction: $U_{norm} < 0$
  - Set factor to 20
    - Reflecting real-world importance of recall for MGI

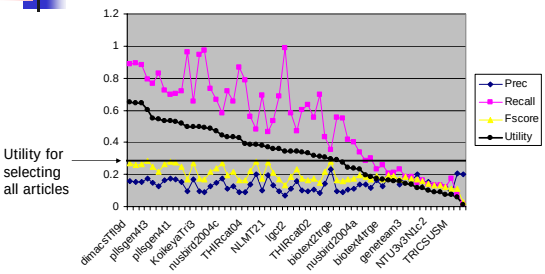
22

Annotation subtasks measurements

- Primary – F-score for tuples (document, gene, GO category, ± evidence code)
  - Recall = correct tuples / all tuples
  - Precision = correct tuples / nominated tuples
  - F = 2 * recall * precision / (recall + precision)

23

Triage subtask results



U_{norm}: n = 59, Max = .6512, Median = .3425, Min = .1114

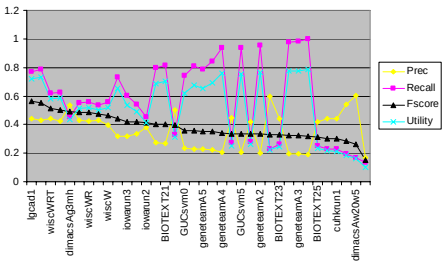
24

Triage subtask analysis

- A variety of features, classifiers, etc.
  - Top results involved MeSH terms from MEDLINE some way
- Of note, Rutgers did a run using solely the MeSH term *Mice* that outperformed all but their top run
  - U_{norm} = .6404, recall = .8929, precision = .1502, F-score = .2572
- Interpretations
  - MGI data is bad – some concerns about MGI data quality
  - Our methods are bad – we do not know or there do not exist better predictive features
  - Our metrics are bad – is factor = 20 appropriate?
  - This is a good and useful finding for MGI

25

Annotation subtask results



F-Score: n = 36, Max = .5611, Median = .3556, Min = .1492
(Also: three runs in annotation plus evidence codes subtask)

26

Annotation subtask analysis

- Best approaches used combinations of
  - Named entity recognition for genes
  - Recognizing document structure of scientific papers
    - e.g., abstract, introduction, methods, results, conclusions
  - Classifiers to learn associations between genes and other features localized to parts of documents with GO codes

27

Future directions

- Continuation of track until (at least!) 2008, thanks to NSF grant
- Aim to develop enduring test collections from 2004 track data
  - Issue: update data?
- Future goals (from 2003 roadmap) include
  - Full-text retrieval
    - Note: Hard to procure; MEDLINE still entry point for most to literature
  - Interactive user experiments
  - Broader types of users, information needs, tasks

28