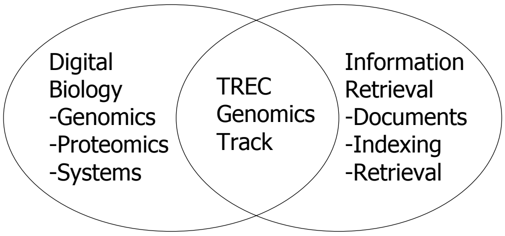
TREC Genomics Track Plenary

William Hersh
Track Chair
Oregon Health & Science University
hersh@ohsu.edu
http://medir.ohsu.edu/~genomics

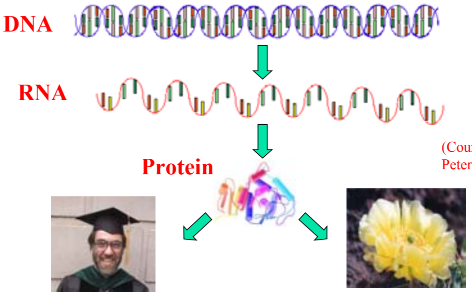
Overview

- Introductory comments
- Track history
- 2003 track
  - Primary task
  - Secondary task
- Future directions

At the intersection of digital biology and IR



The "central dogma" of biology



(Courtesy of
Peter Smookler)

Overview of this session

- 11-11:40 – Overview (Hersh)
- 11:40-12 – University of California, Berkeley (Hearst)
- 12-12:20 – Erasmus Medical Center (Weeber)
- 12:20-12:40 – National Library of Medicine, University of Maryland (Aronson)
- 12:40-1 – University of Waterloo (Clarke)
- 1 – LUNCH!

TREC Genomics Track history

- 2000 and before – Stated desire among TREC participants for track using more structured data (as opposed to just documents)
- 2001 – Suggestion to consider genomics data
- 2002 – Pre-track: Web survey, email list, and organization of workshops
- 2003 – First year of track, development and funding of roadmap for future years

Pre-track Web survey

- Set up as Web survey open to all
- Publicized via many email lists
- Open-ended, asking about interests in data sources and user tasks
- Carried out in spring, 2002
- Obtained about 80 replies

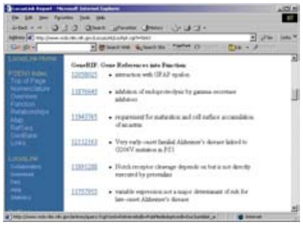
Survey results

- Diverse interests in information retrieval/ extraction tasks, but clustered around three areas
  - Extraction of knowledge from databases
  - Automating the annotation of genes and proteins
  - Retrieval across heterogeneous databases
- Most respondents were interested in using public databases, mainly those from NLM/NCBI

Follow-on to survey

- Workshops at
  - Joint Conference on Digital Libraries 2002
  - TREC 2002
  - Pacific Symposium on Biocomputing 2003
- Each workshop continued refinement of discussion, with general consensus emerging at PSB 2003
  - Tasks defined around use of GeneRIFs for "gold standard" of document relevance and gene function, an assumption not always warranted

What is a GeneRIF (gene reference into function)?



GeneRIF: Gene References into Function

- [illegible] • interaction with GFAP epsilon
- [illegible] • inhibition of endoproteolysis by gamma-secretase inhibitors
- [illegible] • requirement for maturation and cell surface accumulation of nicastrin
- [illegible] • Very early-onset familial Alzheimer's disease linked to G206V mutation in PS1
- [illegible] • Notch receptor cleavage depends on but is not directly executed by presenilins
- [illegible] • variable expression not a major determinant of risk for late-onset Alzheimer's disease

A statement about gene function as described by a publication in MEDLINE (with link to PubMed), assigned systematically since April, 2002 (Mitchell, AMIA, 2003).

TREC 2003 Genomics Track

- Constrained by lack of resources but partially overcome by great enthusiasm
- Primary task – ad hoc document retrieval
  - A reasonable starting task, driven by resource constraints for relevance judgments, GeneRIFs
- Secondary task – identifying text of GeneRIF
  - A combination of extraction and summarization

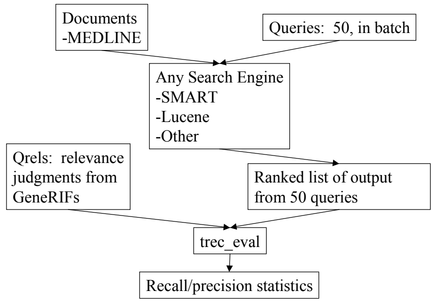
Participation

- Primary track
  - 49 runs from 25 groups
- Secondary track only
  - 24 runs from 15 groups
- Number of groups
  - Initially signed up – 56
  - Primary track only – 16
  - Secondary track only – 5
  - Both tracks – 9

Primary task

- Ad hoc document retrieval task applied to genomics
  - Use case – researcher or graduate students exploring a new domain
  - Metric of performance – MAP
  - Topics – task of finding articles in MEDLINE that provided information on function of a gene

Overview of primary task



Documents -MEDLINE

Queries: 50, in batch

Any Search Engine -SMART -Lucene -Other

Qrels: relevance judgments from GeneRIFs

Ranked list of output from 50 queries

trec_eval

Recall/precision statistics

Content

- Chosen based on use of GeneRIFs as pseudorelevance judgments
- Used a subset of MEDLINE from time period after GeneRIFs became routinely assigned
  - 525,938 records from 4/1/2002 to 4/1/2003

Topics

- Consisted of gene name and instruction to find MEDLINE references about function of gene
  - For gene X, find all MEDLINE references that focus on the basic biology of the gene or its protein products from the designated organism. Basic biology includes isolation, structure, genetics and function of genes/proteins in normal and disease states.

Topics (cont.)

- Used a diversity of gene names
  - Genes with small (i.e., 3) to large (i.e., dozens) number of GeneRIFs
  - Genes represented and not represented in MeSH (former easier to search on)
  - Four different organisms: human, rat, mouse, fruit fly

Topic example

Topic 35: interleukin 3 (colony-stimulating factor, multiple) gene, Locus Link ID 3562, *Homo sapiens*

OFFICIAL_GENE_NAME	interleukin 3 (colony-stimulating factor, multiple)
OFFICIAL_SYMBOL	IL3
ALIAS_SYMBOL	IL-3
ALIAS_SYMBOL	MCGF
ALIAS_SYMBOL	MULTI-CSF
PREFERRED_PRODUCT	interleukin 3 precursor
PRODUCT	interleukin 3 precursor
ALIAS_PROT	mast-cell growth factor
ALIAS_PROT	P-cell stimulating factor
ALIAS_PROT	hematopoietic growth factor
ALIAS_PROT	multilineage-colony-stimulating factor

Relevance judgments

- Based on GeneRIFs
- This means GeneRIFs could be used in searching (but rest of LocusLink could)
- Also carried out analysis to determine accuracy of GeneRIFs as indicators of relevance

Results with test data for 49 official runs

Organization or designation	MAP	Relevant @ 10	Relevant @ 20
National Library of Medicine #1	0.4165	3.16	4.84
National Library of Medicine #2	0.3994	3.20	4.56
National Research Council #1	0.3941	2.94	4.38
University of California Berkeley	0.3912	3.06	4.46
National Research Council #2	0.3771	2.76	4.36
Median	0.2001	1.50	2.44
Gene names	0.1372	1.18	0.88
Lowest	0.0271	0.22	0.60

Listen to follow-on talks, read papers, and visit posters for details of what groups did to obtain their results.

Summary of results from a topic

Topic 35: interleukin 3 (colony-stimulating factor, multiple) gene, Locus Link ID 3562, *Homo sapiens*

	Best	Median	Worst
MAP	0.4136	0.0647	0
Relevant @ 10	4	1	0
Relevant @ 20	6	1	0

What did groups do?

- National Library of Medicine
  - Research group not involved in library operations
  - Used biomedical domain-specific search engine used for ClinicalTrials.gov database for best run (NLMUMDSE)
  - Added mapping into controlled vocabulary and use of linguistic term collocations (NLMUMDSRB) that slightly degraded performance
- Other top-ranking groups also used a variety of domain-specific approaches

How did “standard” IR approaches do?

- Highest results from non-domain specific approaches came from University of Waterloo
- SMART: University of Neuchâtel found best results with Okapi weighting, pivoted normalization, and query expansion, with results near median
- Language modeling: UIUC used variant of language modeling and also performed near median
- Phrases: OHSU used mapping to phrases (and other approaches), scored below mean

How good were GeneRIFs for relevance judgments?

- We assessed with topics looking at:
  - False positives: Are GeneRIFs truly relevant?
  - False negatives: Are relevant documents not designated as GeneRIFs?
- Training topics: For 10 topics, looked at all GeneRIFs and top 20 documents retrieved by best OHSU run
- Test topics: Repeated analysis for all 50
- All assessments done by Dr. Sarah Corley, an OHSU informatics graduate student

Assessing relevance of GeneRIFs

- For 10 training topics
  - All GeneRIFs relevant
- For 50 test topics
  - Virtually all GeneRIFs represented relevant documents (97.3%)

Relevance analysis summary

Category	**10 Training Topics**	**50 Test Topics**
GeneRIF and relevant	10.5%	12.7%
GeneRIF and relevant in another species	0%	0.2%
Not a GeneRIF and relevant	42.5%	41.2%
Not a GeneRIF and relevant in another species	12.5%	36.3%
Not a GeneRIF and not relevant	35.0%	9.2%

Conclusions from relevance analysis

- GeneRIFs are very accurate indicators of document relevance but significantly incomplete
- The number of incomplete GeneRIFs is highly variable across genes
- The problem of documents which are relevant about the gene but in another species is also significant
  - Such documents not necessarily not relevant to real users!

Secondary task

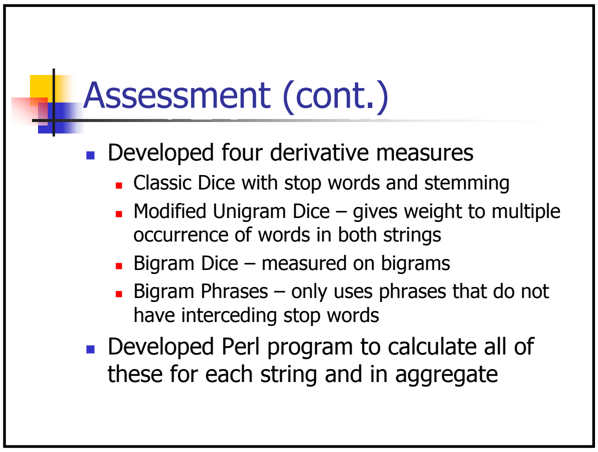
- Goal was to nominate GeneRIF text
- More exploratory since many unanswered questions about quality, consistency, etc. of GeneRIF text
- Some preliminary work by Jim Mork and Lan Aronson showed
  - 95% of snippets came from title and abstract
  - 42% were direct cut and paste from abstract
  - 25% contained significant runs of words

Secondary task – data

- Chose 139 GeneRIFs where we could obtain full-text of documents from publishers who have worked with Highwire Press to allow their content to be used for research
- All GeneRIFs/articles came from five journals (J Biol Chem, J Cell Bio, Nuc Acid Res, Proc NAS, Science) and were published in latter half of 2002

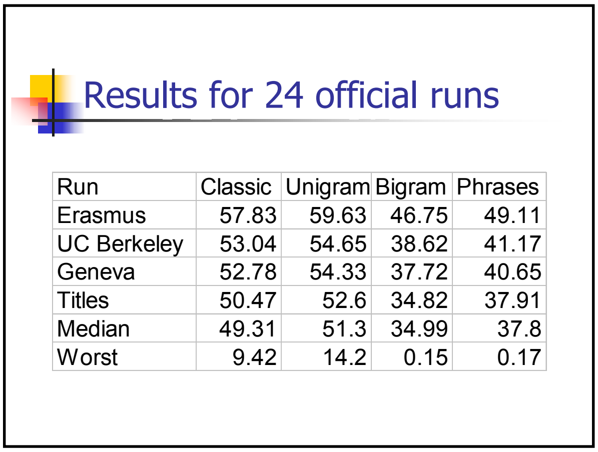
Secondary task – assessment

- Original plan was to use Dice coefficient to measure overlap of GeneRIF and candidate string
  - For two strings A and B,
  - X is the number of words in A
  - Y is the number of words in B
  - Z is number of words occurring in both A and B:
  - Dice (A, B) = (2 * Z)/(X + Y)
- Measure was limited, since does not allow normalization or phrase designation



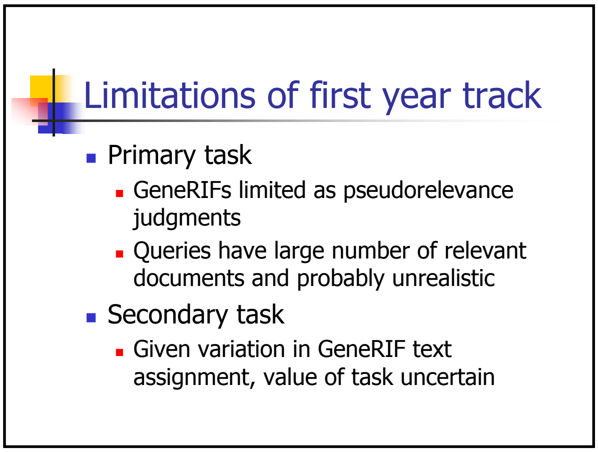
Assessment (cont.)

- Developed four derivative measures
  - Classic Dice with stop words and stemming
  - Modified Unigram Dice – gives weight to multiple occurrence of words in both strings
  - Bigram Dice – measured on bigrams
  - Bigram Phrases – only uses phrases that do not have interceding stop words
- Developed Perl program to calculate all of these for each string and in aggregate



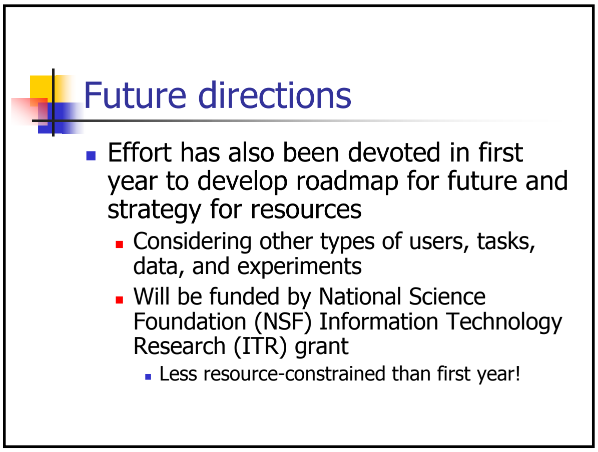
Results for 24 official runs

Run	Classic	Unigram	Bigram	Phrases
Erasmus	57.83	59.63	46.75	49.11
UC Berkeley	53.04	54.65	38.62	41.17
Geneva	52.78	54.33	37.72	40.65
Titles	50.47	52.6	34.82	37.91
Median	49.31	51.3	34.99	37.8
Worst	9.42	14.2	0.15	0.17



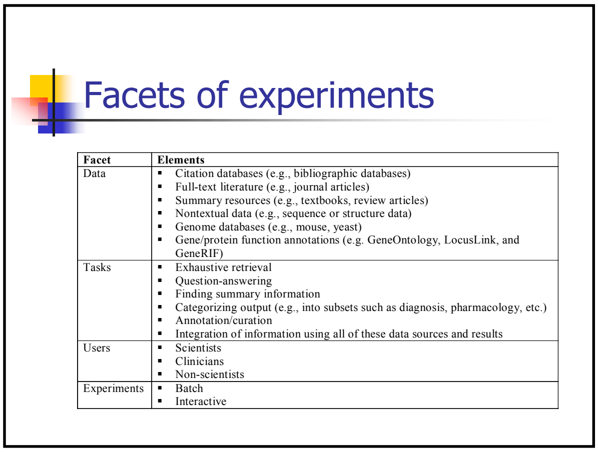
Limitations of first year track

- Primary task
  - GeneRIFs limited as pseudorelevance judgments
  - Queries have large number of relevant documents and probably unrealistic
- Secondary task
  - Given variation in GeneRIF text assignment, value of task uncertain



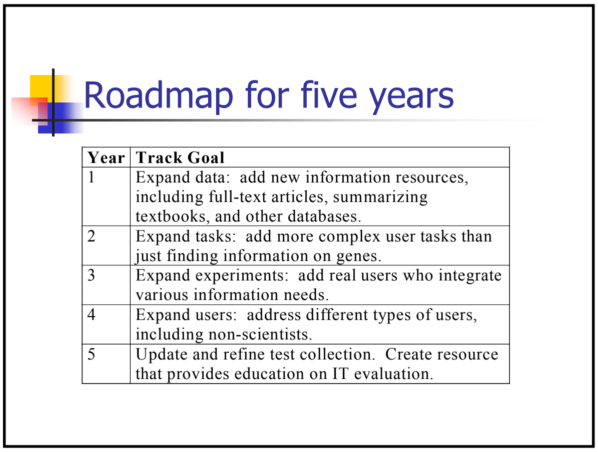
Future directions

- Effort has also been devoted in first year to develop roadmap for future and strategy for resources
  - Considering other types of users, tasks, data, and experiments
  - Will be funded by National Science Foundation (NSF) Information Technology Research (ITR) grant
    - Less resource-constrained than first year!



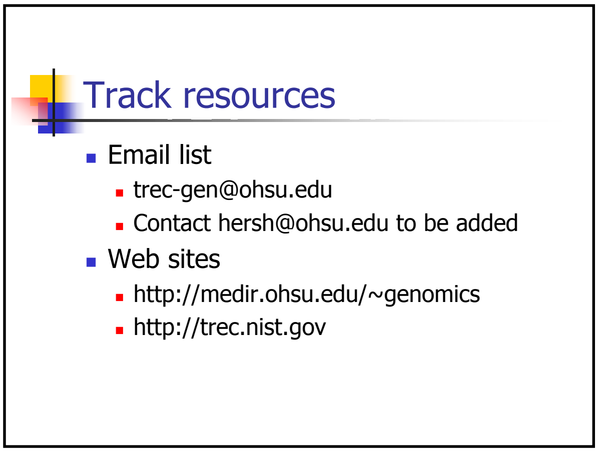
Facets of experiments

Facet	Elements
Data	▪ Citation databases (e.g., bibliographic databases) ▪ Full-text literature (e.g., journal articles) ▪ Summary resources (e.g., textbooks, review articles) ▪ Nontextual data (e.g., sequence or structure data) ▪ Genome databases (e.g., mouse, yeast) ▪ Gene/protein function annotations (e.g. GeneOntology, LocusLink, and GeneRIF)
Tasks	▪ Exhaustive retrieval ▪ Question-answering ▪ Finding summary information ▪ Categorizing output (e.g., into subsets such as diagnosis, pharmacology, etc.) ▪ Annotation/curation ▪ Integration of information using all of these data sources and results
Users	▪ Scientists ▪ Clinicians ▪ Non-scientists
Experiments	▪ Batch ▪ Interactive



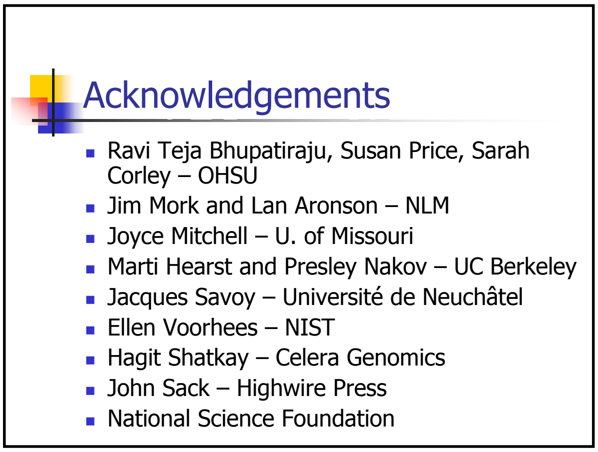
Roadmap for five years

Year	Track Goal
1	Expand data: add new information resources, including full-text articles, summarizing textbooks, and other databases.
2	Expand tasks: add more complex user tasks than just finding information on genes.
3	Expand experiments: add real users who integrate various information needs.
4	Expand users: address different types of users, including non-scientists.
5	Update and refine test collection. Create resource that provides education on IT evaluation.



Track resources

- Email list
  - trec-gen@ohsu.edu
  - Contact hersh@ohsu.edu to be added
- Web sites
  - http://medir.ohsu.edu/~genomics
  - http://trec.nist.gov



Acknowledgements

- Ravi Teja Bhupatiraju, Susan Price, Sarah Corley – OHSU
- Jim Mork and Lan Aronson – NLM
- Joyce Mitchell – U. of Missouri
- Marti Hearst and Presley Nakov – UC Berkeley
- Jacques Savoy – Université de Neuchâtel
- Ellen Voorhees – NIST
- Hagit Shatkay – Celera Genomics
- John Sack – Highwire Press
- National Science Foundation