TREC Genomics Pre-Track Workshop Report

August 8, 2002

William Hersh, Pre-Track Chair, hersh@ohsu.edu

The workshop was held on Thursday, July 18, 2002 at the Joint Conference on Digital Libraries. We had a productive day of presentations and discussion. We emerged with a plan to move forward on the pre-track, a draft of which is presented in this document. This document will be circulated on the listserv (trec-gen@ohsu.edu) and other places.

We are also developing a Web page for the pre-track at the URL:

- http://medir.ohsu.edu/~genomics

At the workshop, it was evident that this embryonic community has a diverse array of research interests and capabilities. We will therefore need to continue working towards a common task and set of resources and evaluation metrics to use for those tasks. Our first step in defining the tasks, resources, and metrics will be to collect a sample task from each individual/group interested in participating, and collating them to determine commonality. *We ask that those planning to participate in this process to submit your sample task via email to the Pre-Track Chair (hersh@ohsu.edu) no later than August 30, 2002*.

The task template will consist of the following:

Input:

- Gene(s) or Protein(s)

Task, including:

- Description
- Motivating context
- Sub-tasks within overall task
- Resources to be used
- Reasons why task is challenging
- Metrics for measuring success and what resources will be required for their use

The resources to be used might include but are not limited to:

- MEDLINE
- OMIM
- GenBank (including dbSNP)
- LocusLink
- Highwire Press - full text of 300+ journal articles (copyright issues?)
- Gene Ontology (GO)
- Enzyme Classification System
- UMLS (users will need free license from NLM)
- HUGO Gene Nomenclature Committee (HGNC) codes
- SGD (Saccharomyces Genome Database, with GO codes for gene names and articles)

- Flybase
- SWISSPROT - links protein sequences to each other and literature (copyright issues?)

Output(s), including metrics:
- Winnowed articles
- Potential GO codes
- Interactions with other genes
- Equivalent genes in other organisms

We also articulated a timeline:
- August - get one fully instantiated task from each group that desires to participate
- September - collate tasks into database or spreadsheet
- October - analyze tasks
- November - present analysis and meet at TREC (probably 11/18)
- January - meet at PSB 2003 to identify sources, tasks, assessors, funding, etc.
- Spring - distribute data (or how to get it) plus sample tasks
- Summer - distribute tasks and run experiments
- Fall, 2003 - assessment of results
- November, 2003 - present first track results at TREC 2003 meeting

Here is an example of an instantiated task template:

Gene:
- PSEN1

Tasks
- Description - This gene is known to be associated with some forms of early-onset Alzheimer's Disease. Has this gene been discovered in non-humans and, if so, has it been associated with any diseases?
- Motivating context - Searching for an animal model where not only the same mutation(s) in the homolog gene exist but there must also be a corresponding disease, for the purpose of development or testing of novel therapeutic agents in preclinical trials.
- Sub-tasks within overall task
    - Has the gene been discovered in non-humans?
    - Has the gene been associated with diseases in non-humans?
- Resources to be used - MEDLINE, OMIM, GenBank, LocusLink, GO, SGD, Flybase
- Reasons by task is challenging - there are aliases for PSEN1, they include: AD3, PS1, PS-1, S182, Presenilin 1.
- Metrics for measuring success and what resources will be required for their use
    - Correct answer of question with supporting evidence
    - Documents describing association with diseases (recall, precision)

Output:
- MEDLINE citations describing the homolog genes
- MEDLINE citations describing associated diseases