A MITRE TECHNICAL REPORT

Patterns of Pathogenesis

MITRE Sponsored Research Final Report

September 2006

Lynette Hirschman
Alexander Morgan
Marc Colosimo

Sponsor: MITRE Sponsored
Dept. No.: G063
Derived By:

Contract No.:
Project No.:
Downgrade To:
Declassify On:

MITRE
Center for Integrated Intelligence Systems
Bedford, Massachusetts

Abstract

This MSR has focused on the development of tools to analyze infectious pathogens. The rationale is that many pathogens share a set of functions that allow them to invade a host cell, evade the host cell defenses, multiply inside the host cell, and eventually escape both the cell and the host organism to spread infection. Our goal has been to bring to bear the rich set of bioinformatics resources that are becoming available, from gene sequences to knowledge embedded in the biological literature, in order to understand these "virulence factors." We have focused on: 1) identifying relevant datasets and resources; 2) developing a pipeline for analysis of experimental data; and 3) developing flexible tools to integrate information from the biomedical literature. A deeper understanding of virulence mechanisms will make it possible to create improved disease models, to identify countermeasures, and to speed up the "bug-to-drug" pipeline. Our accomplishments include the creation of an international challenge evaluation for text mining in biology (BioCreAtIvE); the creation of an international community focused on text mining tools to support for curation of biological databases; support to DARPA to pitch a BioOntologies program; the award of a grant from NSF; and the publication of 20 peer reviewed papers and book chapters.

Table of Contents

List of Tables

List of Figures

1 Introduction

New diseases are constantly emerging and concern with biosecurity looms large in defense and public health planning. The urgency of the problem has been illustrated by the 2001 anthrax attacks, the impact of the SARS outbreak in 2002, and the current threat of H5N1 avian influenza that could mutate into a human transmissible form with devastating pandemic potential. The rapid elucidation of the mechanisms (virulence factors) by which these microbes cause harm is key to an informed, effective, rapid response.

There has been significant government investment in improving the state of knowledge and national preparedness. The Defense Science Board's 2001 Summer Study describes a "Pathogen to Hit" program to move toward "Bug to Drug in 24 hours." The NIH's NIAID (National Institute of Allergy and Infectious Diseases) has undertaken the establishment of eight Bioinformatics Resource Centers, focused on sequencing of CDC Class A, B, and C pathogens. The FDA Critical Path program is heavily focused on development of new biomarkers, as well as ways to speed drug development.

An understanding of virulence mechanisms lies at the heart of this much larger research activity. This MIITRE-Sponsored Research (MSR) project has been focused on host-pathogen interaction. Many pathogens share a set of functions that allow them to invade a host cell, evade the host cell defenses, multiply inside the host cell and eventually escape both the cell and the host organism to spread infection. Our goal has been to bring to bear the rich set of bioinformatics resources that are becoming available, from gene sequences to knowledge embedded in the biological literature, in order to understand these "virulence factors." We have focused on: 1) identifying relevant datasets and resources; 2) developing a pipeline for analysis of experimental data; and 3) developing flexible tools to integrate information from the biomedical literature. Our approach was to create a pipeline to analyze data from a high-throughput experiment, applying bioinformatics techniques to add meta-data to clusters of genes. This pipeline could then provide connections to support biologically based interpretations of the observations; our focus for the MSR was based on an experiment looking at host response to mouse-adapted 1918 influenza. This work allowed us to identify needed resources and, in particular, the need for more flexible tools to extract critical information from the biomedical literature.

This has led to identification of major gaps in:

- Biological knowledge in computable form (functional annotation of genes and gene products, pathways, and virulence mechanisms);
- Computable access to the information contained in the biomedical literature, which serves as the major repository for biological knowledge;
- Tools for the analysis of data, to permit integration of experimental data, e.g., from high throughput experiments, integrated with bioinformatics data and information

from the literature, including tools for the management of complex biological workflows.

The accomplishments under the MSR are:

- *An initial pipeline for the analysis of high throughput micro-array data* that has been applied to internal experiments.
- *Establishment of BioCreAtIvE:* Critical Assessment of Information Extraction in Biology, the first international challenge evaluation for text mining in biology that evaluates text mining applied to creating, maintaining and accessing information stored in the biomedical "bibliome," in association with NCBI, the Spanish National Center for Cancer Research, as well as the MINT and IntAct protein interaction databases.

2 Programmatics

2.1 Objective

The objective of this work has been to develop tools to analyze infectious pathogens, with a focus on biosecurity from two perspectives: First, the analyst's perspective, with a goal of understanding virulence mechanisms better, e.g., why was 1918 influenza so lethal? What would it take for H5N1 to become human transmissible? And second, the biotechnology perspective, focused on improving bug-to-drug pipeline, e.g., what are pathways involved in pathogenicity?

2.2 Approach

Overall, our approach has focused on identification and integration of resources, including data (both structured and free text data) and tools to access, clean, map, and integrate across diverse data types and data sources. These can be divided into three major efforts, which roughly track the three years of the project:

- Assembling data resources;
- Experimental data analysis;
- Integration of data from multiple sources, including the literature.

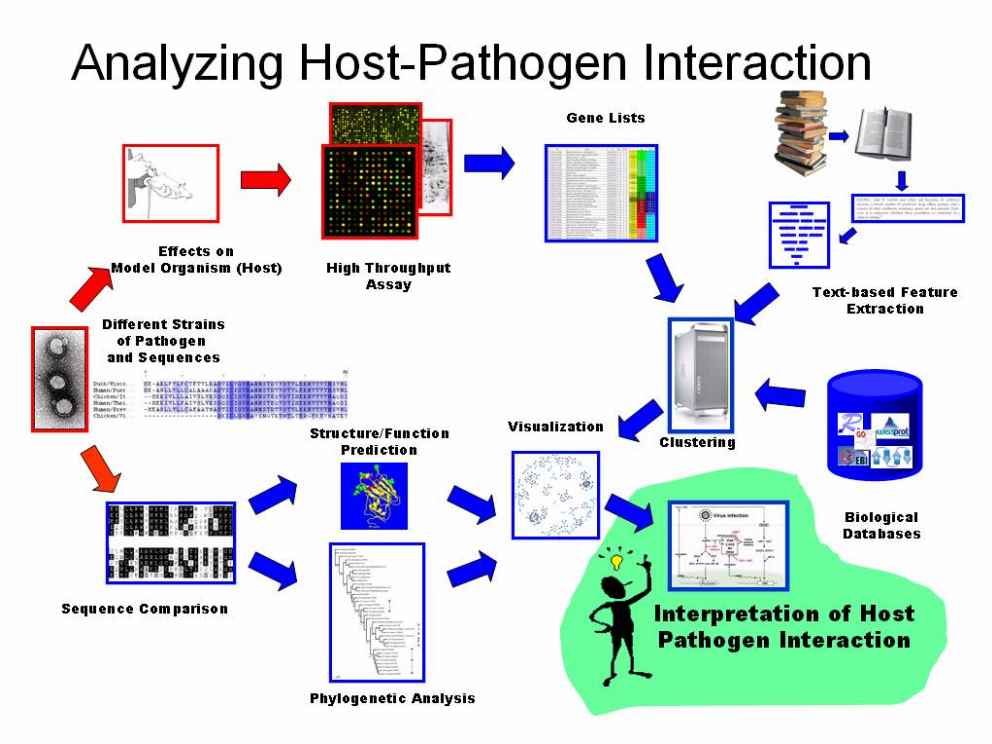
The work that we have pursued is integral to turning biological data collection and integration activities into a repeatable computer-supported analytic process (**Error! Reference source not found.**). Our original work looked at the virulence factors of the individual pathogens, while our later work has been addressing the understanding the infection as a perturbation of host (e.g., human or mouse) physiology. These are, of course, complementary approaches, but the former is perhaps more appropriate to most bacteria, and the latter to viruses, which depend so much on the molecular biology of the infected host. In support of these goals, we have examined methods for the improved interpretation of the massive number of data points resulting from microarray experiments, in the context of simultaneously extensive, but incomplete, knowledge of the functional relationships between the genes and proteins of the host organism. We have also explored using our expertise in biomedical text processing to augment source data on the host genes and proteins. The goal was to bring us one step closer to explaining puzzles, such as how minute differences in the protein products of an influenza virus affect virulence and become the determinants of life or death of the host organism.

Figure 1 shows how multiple threads of evidence can be brought to bear in understanding host-pathogen interaction and the mechanisms of pathogenicity. The upper thread traces an experiment from exposure of host organism to multiple strains of a pathogen, to a micro-array measuring expression of gene products under different conditions, to an interpretation of the results, drawing on information from biological databases and the literature. The bottom thread traces an approach (now being pursued under the Genomics for Bioforensics MSR), looking at genetic changes that result in different gene products (proteins) with different properties; these strains evolve over time, escape immune response and cause new outbreaks. The information

from these multiple sources needs to be brought together, integrated and visualized, in order to create a biological interpretation of the data (lower right-hand corner), which, in turn, leads to new hypotheses.

The primary technical areas of bioinformatics which we have pursued can themselves be divided into two main areas. The first is the area of biomedical text mining, an area in which MITRE has become a well established authority. The second area is the integration of heterogeneous data sources (e.g., sequence data, expression data, physiological function annotation, etc.) to support analysis. Although these closely track more general analytic tasks and areas of existing MITRE expertise, the particulars of the domain require specialization, and it is hoped that future work at MITRE will both allow us to increase our corporate expertise in biomedical information analysis and to apply knowledge gained in these areas to support other informatics tasks

Figure 1 Analysis of Host-Pathogen Interaction



3 Toxin and Virulence Factor Database at Los Alamos

Bioinformatics is an information science that helps to both organize and arrange the biological data that biomedical researchers provide, to form a gestalt from which inferences can be drawn that would not have been realized otherwise. It is an attempt to take advantage of the forest of results formed from all the trees and bushes lovingly planted as individual experiments. Knowing this and wanting to use our computational techniques to help explore the complexities of pathogens and their complex interactions with their hosts, we initially partnered with a data provider.

In 2004, a group headed by Murray Wolinksy at Los Alamos National Labs was developing the TVFac, the Toxin and Virulence Factor Database (http://www.tvfac.lanl.gov/). Los Alamos was building on extensive experience sequencing and annotating (at the level of gene finding) microbial genomes and hosting pathogenic sequence information (GenBank, HIV Sequence Database, Influenza Database, etc.). They had developed a series of guidelines and a hierarchy of virulence factors and had begun an effort to systematically classify and annotate the known (published in the research literature) virulence factors for all the significant human pathogens. This seemed an excellent opportunity for MITRE to leverage past work doing biomedical text mining to aid database curators/annotators in their efforts, while providing opportunities to use the data collected by Los Alamos to analyze the relationships between virulence mechanism and host interactions.

Our interest in supporting the semi-automated curation of the Toxin and Virulence Factor database motivated us to try two new things: a massive integration of biomedical data, and the development of a new text processing framework that allowed us to rapidly adjust our features, annotations, and data structures. Our previous work had exposed us to some of the main model organism databases (FlyBase, Mouse Genome Informatics, Saccharomyces Genome Database), but this new work required us to become much more familiar with resources we had only worked with indirectly (NCBI, UniProt, GeneOntology). Our previous work had also been with highly structured and richly populated model organism databases. In contrast, the information about pathogens and the primary host of interest (humans) was spread across several different resources and not well integrated. It was a considerable effort to process and integrate the material distributed in these different sources, often dealing with significant issues of data error and incompleteness (errors in file formats, typos for identifiers, orphan concepts, etc.). Some of the mapping and integration code extensions we made were returned to the research community (e.g., contributions to the BioPython Open Source project).

The other principle area in which we developed new infrastructure was in the areas of our basic text processing and machine learning infrastructure. We knew we would need to combine text based features, genomic sequence based features, and richer annotation types together in a flexible modular framework. We developed a basic framework involving a database management system to store our basic features (including text), which we wrapped in intermediate processing to produce a simple vector model to provide input to a library of

machine learning systems (e.g., WEKA, SVMlight MALLET, etc.) to allow for experimentation. At the same time, we created a simple annotation framework using active webpages (PHP) to interface with our database and processing system. All this made for a highly flexible system for experimentation.

4 Biological Databases

Bioinformatics is an information science and this project has been heavily focused on data mining, particularly across heterogeneous data types. At the beginning of the project, we had to spend considerable time collecting data from a variety of sources and providing linkages among the different data source to create a relational database. This was in support of the approach described in the previous section which made heavy use of active webpages (PHP) to create annotation forms and to report views of our compiled data stored in a SQL database. The wide variety of data sources we collected is summarized in Table 4. As with any data integration effort, quite a bit of time was spent with fixing minor errors in the data sets (bad file formats, etc.): there is no systematic solution to such issues.

One of the major problems in bioinformatics continues to be data sparsity. It may seem counter-intuitive that thousands of annotations for genes constitute a sparse dataset, but when used to understand a microarray which has tens of thousands of genes on it, the gaps in the computationally accessible data become apparent. For example, in the experiments with the mice infected with recombinant 1918 influenza that we describe in the following section, of the genes on the microarray, only 35% had any sort of Gene Ontology annotation at all (and many of these were very incomplete even when they had any annotation). The fact that over twice that number, 76% of the genes on the chip, had links to journal articles in MEDLINE motivated our continued work in text mining. For details refer to Section 5.

In general, the mappings between databases are woefully incomplete. This causes many problems when trying to combine diverse data types. For example, the Gene Ontology annotations made for humans use UniProt identifiers. PIR (Protein Information Resource) has done much of the very difficult job of linking those to EntrezGene identifiers (see Table 4). However, it is only 90% complete for those proteins annotated with GO concepts (and less complete for less well studied proteins). That means that trying to link a chromosomal location or particular gene to a functionally annotated protein can't be done 10% of the time for humans, one of the most widely studied organisms.

Table 1 Human Proteins Mapped to EntrezGene

UniProt Portion	Annotated by EBI		Mapped to EntrezGene
Swiss-Prot (7,772)	6970	79%	Y
	802	9%	N
TrEMBL (1,093)	965	11%	Y
	128	1%	N

During the course of this project, we tried not to be merely passive consumers of biological data, but to make strong connections with the data providers and with data standards efforts. We have maintained contacts and have had numerous meetings with data providers from e.g., MGI (mouse genome), EBI (European Bioinformatics Institute, responsible for hosting UniProt and doing human GO annotation), PIR (Protein Information Resource at Georgetown), The Immune Epitope Database, and many others. We have been involved with the Gene Ontology Scientific Advisory Board (L. Hirschman has served on this for three years), Semantic Web for Life Sciences, IEEE Bioinformatics Standard, BioPAX, and other groups.

As this project progressed over its three-year life, the organization and quantity of the publicly available data has increased tremendously, particularly at NCBI/NLM. Numerous groups have worked to provide better access and much of the work we did organizing and linking data sources has been surpassed by domain specialists. For example, the Resourcerer data we used for to map the mouse microarray data to gene identifiers improved significantly on the linkages we developed for the oligos on the chip mapped to EntrezGene identifiers. We have been able to contribute in a small way to tools such as BioPython, which provides functions to wrap the emerging API's for accessing many of the key resources in the python scripting language. As these types of bioinformatics experiments move from bleeding edge to mainstream, the accessibility, organization, currency, and relevancy of the data will continue to improve, making this kind of integration much easier.

5 Functional Microarray Analysis

An RNA microarray is an experimental apparatus, which can be used for a number of different experiments, in particular the measure of the relative expression of thousands of different genes simultaneously. However, the experimental methods themselves are only a small part of expression array analysis; as in other high throughput experiments, there are the experimental techniques; statistical analysis for data cleaning and significance testing; the development of biomarkers; and the analysis of these results in a physiological context. There continues to be research in how to improve all stages of this process. However, the latter aspect, namely relation between the expression array results and the physiological context, is only starting to be explored. This latter aspect is particularly important in the analysis of diseases, because the symptoms of

disease are a perturbation in the normal system, and oftentimes disease processes are intimately tied up to the complex interactions and interrelationships in molecular physiology of the affected organism.

A typical result from a high-throughput experiment is a list of genes or proteins that are differentially regulated under certain conditions. Researchers' ability to run such experiments and collect the raw data has now greatly exceeded their ability to analyze the results in a timely fashion. The challenge is to interpret this list of genes in terms of possible mechanisms that can, for example, explain which sets of genes are co-regulated or are interconnected in pathways. This requires providing sufficient information about the function of individual genes or gene products and their connections in pathways, in order to develop an explanation of the underlying biological processes involved.

This is a typical bioinformatics problem, and also a Semantic Web challenge; it requires the integration of many heterogeneous data resources, such as model organism databases, pathway databases, protein function databases, and (at least ideally) information contained in the literature. These are coupled into a complex workflow using available bioinformatics tools – a process which is time-consuming and requires significant maintenance to obtain reproducible results, as discussed in the previous sections.

There are now standards for the deposition of high-throughput data sets, such as MAGE[1] as well as meta-data standards (MIAME or Minimal Information about Microarray Experiments). The goal of these standards is to permit the capture and sharing of the raw datasets with no error inducing reformatting. These repositories enable "reannotation," making it possible to use new information and new tools that become available over time to reannotate older data sets. In one such reannotation exercise,[2] a set of raw microarray data

[1] Micro-array and Gene Expression: http://www.mged.org/Workgroups/MAGE/mage.html.

[2] Kash, J.C., C.F. Basler, A. Garcia-Sastre, V. Carter, R. Billharz, D.E. Swayne, R.M. Przygodzki, J.K. Taubenberger, M.G. Katze, and T.M. Tumpey, *Global host immune response: pathogenesis and transcriptional profiling of type A influenza viruses expressing the hemagglutinin and neuraminidase genes from the 1918 pandemic virus.* J Virol., 2004. **78**(17): p. 9499-511.

Table 2 Summary of Data Resources

Database	Data	Issues
GenBank	Nucleic acid sequence data; enormous	Highly redundant; meta data incomplete and full of errors; hard to identify association with a particular organism, protein, etc.
EntrezGene	Unique identifiers for genes; organized by organism	Only a fraction of all sequences are linked with Entrez numbers
MeSH	Controlled vocabulary of coding terms to mark concepts in bio medical text	Highly medical in focus; large size
MEDLINE	Collection of bibliographic material including abstracts for millions of biomedical research papers; indexed with MeSH concept	Massive size; errors in mappings to MeSH appear; updated constantly
Enzyme Nomenclature (EC)	Controlled vocabulary of enzyme activities organized in a hierarchical structure	
NCBI Taxonomy	Controlled vocabulary and hierarchy of organisms organized around phylogenetic trees	Coverage is poor for many micro-organism; the concept of species starts to break down
UniProt	Merging of previous protein DB resources (SwissProt, TrEMBL, and PIR); lots of metadata on proteins and links to associations with Entrez, PDB (structure), etc	Highly redundant; same proteins have multiple appearances as unique entities; linkages to other indentifiers incomplete
Gene Ontology	Hierarchy of protein attribute concepts (mostly functionally related) designed for comparative genomics	The resolution and coverage of the annotation varies tremendously
BIND	Databse of protein-protein interactions	Data focusses mainly on yeast and has poor coverage in other areas
KEGG	Metabolic pathway	Pathways are 'generic' and not tied to specific organisms; the focus is mainly on the metabolism of small molecules
BioCarta	Pathways, particularly signal transduction	The coverage is very low; data is hard to extract from BioCarta (may be obtained indirectly through DAVID from NIAID)

was identified and downloaded. The goal of the original experiment was to gain insight into virulence mechanisms and immune response by comparing mice infected with different strains of influenza virus; the experiment had been performed in 2002, prior to extensive expansion of the Gene Ontology. The hypothesis of our reannotation experiment was that use of updated GO codes would provide significant new meta-data to assist in interpretation of the experimental results. This experiment has been described in a technical report by Marc Colosimo.[3]

After re-extraction of the sets of differentially regulated genes, the next step was to find information in biological databases, including MGI and the various pathway databases, to support annotation of the genes. Most of the time in this exercise was spent mapping from one representation or terminology into a different terminology, in order to access a different set of biological resources (e.g., Genbank ID to EntrezGene to MGI identifier).

The results illustrate why access to the literature is critical. Of the 6544 sequences on the microarray, 64% (4229) could be associated with MGI identifiers and 35% had GO annotations (2316). However, 76% (4936) of the genes had PubMed references in MGI, although this number includes largely uninformative citations from large scale sequencing experiments. This suggests that even for a well-annotated organism such as mouse, much of the information is either unknown or not yet captured in the associated model organism database. Furthermore, any attempt to enrich the annotation set by "inheriting" annotation from homologous genes/proteins is likely to suffer from inaccuracies, because these mappings are not supported by an experimental evidence.

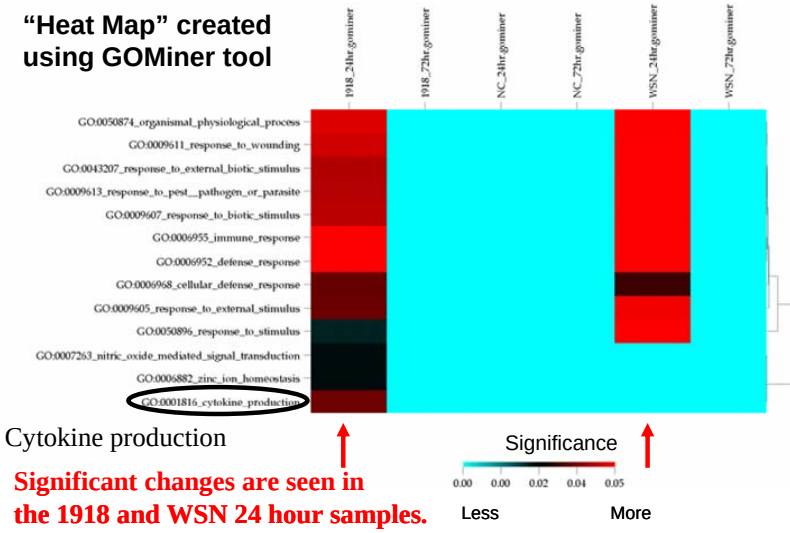
The reannotation experiment allowed us to experiment with a number of tools for functional analysis, listed in Table 3 below. Based on the re-analysis, we were able to confirm, for example, that genes associated with cytokine expression were extensively upregulated for the highly virulent 1918 influenza strain at 24 hours post-exposure, but not for the less virulent strains.

[3] MITRE Technical Report on Functional Analysis of High throughput Experiments (Marc Colosimo, in preparation).

Figure 2 Tools for Functional Analysis

Differences in expression linked to host response pathways

"Heat Map" created using GOMiner tool



Cytokine production

Significant changes are seen in the 1918 and WSN 24 hour samples.

GOMiner associates differentially regulated genes with Gene Ontology codes to provide clues to gene function

Co-regulated genes may also appear in a pathway (e.g., related to immune response)

Goal: understand host response to pathogen in virulent (1918, WSN) and non-virulent strains (NC) over different time periods (columns);
Rows represent different Gene Ontology codes for gene function

Figure 3 Network of Mouse Genes: annotated mouse genes showing pairwise interaction (with ~1800 edges); the 254 nodes in black are those found on the array.

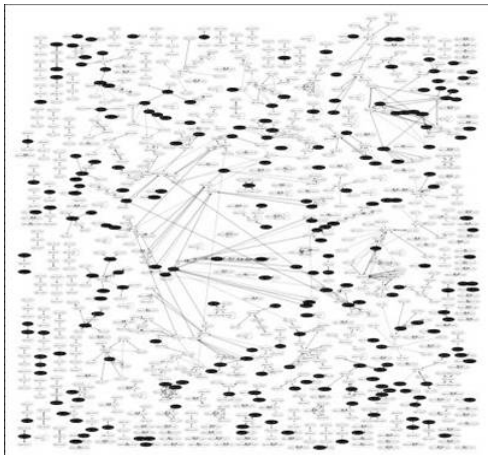


Table 3 Tools for Functional Analysis

Tool	Availability	Overview	Advantages	Disadvantages
GoMiner	Free			
DAVID	Free, Web-based & Download	Takes in a set of common identifiers (Entrez, UniProt, etc) and compares annotations from multiple sources, compares likelihood of shared annotations using simple background model (annotations in source organism)	Excellent source of annotations for the input array; particularly good since it includes pathway DB's including BioCarta (only tool known to use this)	Web interface is rather clunky and prone to failure/stalling out on large datasets; downloadable version is missing many data types;
FatiGO	Free, Web-based	Simple comparison of GO codes with basic distribution of usage	Incredibly simple	Only considers GO annotations
iHOP	Free, Web-based	A text mining utility that does keyword search over annotations and abstracts, can look for colocations	Uses text based features so potentially orthogonal datasets	Does not use common identifiers; impossible to compare long gene lists
Pandora	Free, Web-based	Visualization system to look at probabilities (simple shared annotation model) of shared annotations in input list	Display of results is different from other systems	Some bugs, hard to get
Igenuity PAS	$$$$, Web-based & Download	A full system for visualizing interactions and comparing annotations for input lists of identifiers (consumes most known chipsets or identifier types); proprietary knowledge base extends public knowledge which is also included	A deep proprietary knowledge base, impressive environment, intuitive and useful; free trial	Expensive, subscription service means continued membership to compare new results
Cytoscape	Free, Download	A network/graph visualization system designed for depicting interactions and pathways along with annotations on those pathways	Nice graphic representation full of tools; numerous groups are contributing packages for things like inputting data sets (including text based interaction prediction); numerous graph analysis tools	Primarily a graph visualization system; something of an overhead to get it working; may require java programming to customize to task

6 Rapidly Retargetable Text Mining

The problems described in the preceding sections illustrate that each problem requires special tailoring in terms of the specific types of information that need to be captured. For text mining, this can add significantly to the cost of creating a tailored application. Text mining must overcome the cost/performance barrier by creating modular tools that are easy to adapt to new requirements and new vocabularies through feedback mechanisms that support rapid tailoring and incremental learning. During this MSR, we have investigated this problem space in the context of organizing BioCreAtIvE (Critical Assessment of Information Extraction in Biology), and in the context of building tools to support the (semi-)automated extraction of information from the biomedical literature (see the discussion in the preceding section about information found in the biomedical literature, but not in curated databases). This section describes a number of activities related to text mining and creation and evaluation of tools for automated extraction of information from the literature.

6.1 Analysis of BioCreAtIvE Results

Our first activity during 2004 was to analyze the results from the initial NSF-funded BioCreAtIvE that took place in 2003-2004. BioCreAtIvE focused on two tasks; the first dealt with extraction of gene or protein names from text, and their mapping into standardized gene identifiers for three model organism databases (fly, mouse, yeast); the second task addressed issues of functional annotation, requiring systems to identify specific text passages that supported Gene Ontology annotations for specific proteins, given full text articles.

The first BioCreAtIvE achieved a high level of international participation (27 groups from 10 countries and provided state-of-the-art performance results for a basic task (gene name finding and normalization), where the best systems achieved a balanced 80% precision / recall or better, which potentially makes them suitable for real applications in biology. The results for the advanced task (functional annotation from free text) were significantly lower, demonstrating the current limitations of text-mining approaches where knowledge extrapolation and interpretation are required. In addition, an important contribution of BioCreAtIvE has been the creation and release of training and test data sets for both tasks. Our activities as organizers of BioCreAtIvE led to a number of talks and a special issue of a journal, devoted to BioCreAtIvE (BMC Bioinformatics 2005, 6(Suppl 1). This work was done by Colosimo, Colombe, Hirschman, Morgan and Yeh.

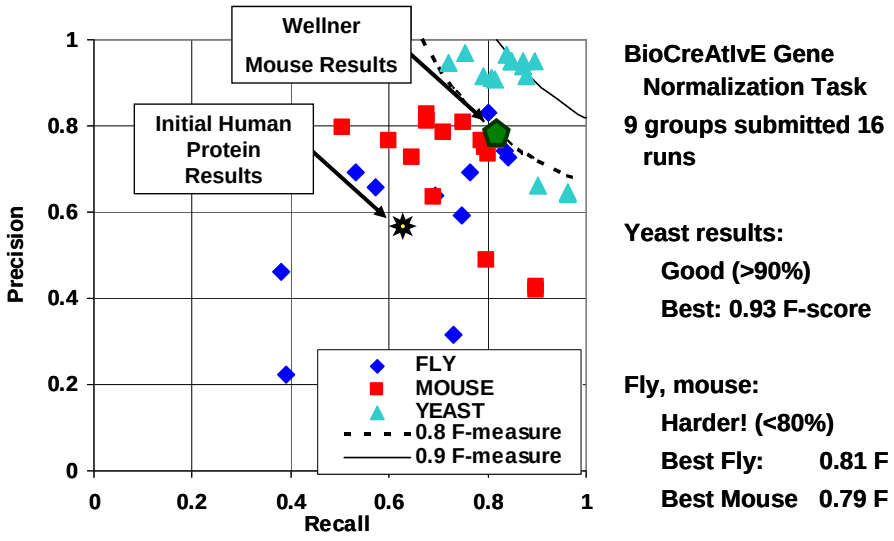
6.2 Retargetable Entity Normalization

Based on lessons learned from the first BioCreAtIvE, Wellner applied the MITRE Carafe toolkit to the problem of entity normalization. A basic set of tasks facing biomedical text processing systems is that of categorizing, identifying and classifying entities within the literature. A key step in this process involves grouping mentions of entities together into equivalence classes that denote some underlying entity. In the biomedical domain, however, we are fortunate to have structured data resources such as databases and ontologies with entries denoting these equivalence

classes. In biomedical text mining, then, this process involves associating mentions of entities with known, *existing* unique identifiers for those entities in databases or ontologies – a process referred to as *normalization*. This ability is required for text processing systems to associate descriptions of concepts in free text with a grounded, organized system of knowledge more readily amenable to machine processing.

Our approach leveraged that taken by the one system at BioCreAtIvE (from U Penn) that was developed to require little domain-specific tailoring. The approach uses a lexicon, coupled with some fuzzy pattern-matching, to tag candidate biological entities, followed by a maximum entropy classifier to associate candidate entities with unique identifiers to generate a list of gene identifiers mentioned in each abstract. The novel approach introduced in (Wellner 2005) is to relabel "noisy" (incompletely annotated) data, in order to obtain better training data, to iteratively improve the model. The results produced by the Carafe based system were comparable to state of the art results from BioCreAtIvE, but using more flexible technology that was later applied to an entirely different domain (see Section 6.5 below).

Figure 4 Results on Gene Normalization Task from BioCreAtIvE compared to Carafe Based Results



6.3 Organization of BioCreAtIvE II

Under the MSR, we wrote a grant to NSF to run a second BioCreAtIvE and to extend the applications to handle complex relations, such as those found in host-pathogen interaction or

complex ecological relations. The NSF award has been funded and will begin on October 2, 2006.

6.4 Creation of a Gene Normalization Data Set

As part of our preparation for BioCreAtIvE, Alex Morgan led the creation of a second gene normalization data set, this time focused on human genes. To create the data set, abstracts were selected from those annotated by EBI's Human GOA group, since this selection was assumed to be enriched in mentions of human genes and gene products. A small group of annotators trained in molecular biology searched through the abstract text (and title), identifying mentions of genes and gene products using UniProt and using the NCBI Gene interface for identifying the corresponding EntrezGene identifier. Inter-annotator agreement was measured at over 90%. A hand annotated training/development set of 281 annotated abstracts was released as training data, and we have prepared another 300 annotated abstracts to be used as blind test data in the evaluation. We also compiled a lexicon for the human EntrezGene identifiers using common gene/protein name sources, which has been released along with the training data. Five thousand abstracts from the GOA annotation set have also been released, along with the EntrezGene identifiers that correspond to the EBI GOA annotations. These have been derived by mapping from the Uniprot to the EntrezGene mapping of PI and may provide useful noisy training data. However, there are a number of limitations with this dataset set since most gene/proteins mentioned are not recorded, and the annotations which were done to UniProt do not completely map into EntrezGene. Participants are requested not to download or use the EBI human GOA annotations on their own.

This work is described in a paper that has been accepted at the Pacific Symposium for BioComputing (Morgan et. al., 2007). The paper describes a set of experiments that were run to validate the data, including estimating the limitations of simple lexical matching, estimates of the quality of the noisy training data, and looking at the biological relationships between genes and proteins which are mentioned together in the same abstracts. This work has provided an additional data point and new insights in understanding issues of task difficulty and domain portability.

6.5 Rapid Retargeting for Medical Data

Based on our successes with rapid retargeting for biological entities, we extended the Carafe-based approach to an anonymization task for clinical (medical) data. This work was the result of a collaboration between teams at MITRE Bedford and the Harvard Center for Biomedical Informatics. We took advantage of the ***AMIA Challenges in Natural Language Processing for Clinical Data,*** which represents the first shared task for the application of natural language processing technology to clinical data. Anonymization of clinical records is a key stepping stone toward the capture of information in clinical records. If systems can reliably identify Protected Health Information (PHI), such as patient name, doctor name, dates and identifiers, such systems also can be used to extract other important information, such as medications or diagnoses. This

latter application aligns with our long-term goal of extracting information from biomedical literature and mapping this information into standard biomedical terminologies or ontologies.

Our approach focused on rapid adaptation of existing toolkits for named entity recognition. We submitted three separate runs, two based on the Carafe[4] toolkit developed at MITRE, and the one using LingPipe[5], a commercial product from Alias-I. These experiments focused on what needed to be done to train the system, how well the system worked "out of the box," whether there was adequate training data, and how much work was needed for additional performance gains. The results on a held-out set of the training data were excellent, with a balanced recall /precision of 0.975/0.986 for a Carafe system tuned to the domain, and a "high recall" system with recall/precision of 0.981/0.974. The MITRE results were the highest reported results for the anonymization task. The paper describing the results was presented at the AMIA November meeting.

[4] http://sourceforge.net/projects/carafe
[5] http://www.alias-i.com/lingpipe

7 Biomedical Ontologies

Under this MSR, we have become heavily involved in activities related to the creation and curation of biomedical databases. In the early stages of the MSR, we worked with Wolinsky and Song at Los Alamos on TVFac (Toxin and Virulence Factor Database, see Section 3). In the course of organizing the first BioCreAtIvE, we worked with FlyBase, Mouse Genome Informatics for the Mouse Genome, and also with the Yeast database, as well as with curators from Gene Ontology annotation team at EBI. One outcome of this involvement is that Hirschman is now serving on the Scientific Advisory Board of the Gene Ontology Consortium.

During 2005, we supported Dr. Sri Kumar from DARPA IPTO to put together a new program pitch for BioOntologies. The starting point was that encoding of biological information into semantically "computable" form would be critical to progress in biology. Hirschman and Colombe organized a workshop that led to a proposal for "DisARM: Disease Agent Rational Modeling" based on using ontologies to capture information about host pathogen interaction to support better disease modeling and prediction, as well as improved drug discovery. More recently, Alex Morgan worked with the Immune Epitope Database, as well as with EntrezGene and UniProt in the course of preparing the BioCreAtIvE Gene Normalization data set.

Curation of knowledge from the published literature is a key source of the information for the biological databases. The curation activity is managed in a curation pipeline consisting of three stages:

1. Management of the curation queue; this involves selection of the literature to be curated, according to some agreed upon set of criteria and priorities;
2. Listing of "curatable" entities (genes, gene products, proteins) in a given paper, linked to their unique identifier;
3. Curation of the list of entities in #2 above, often including annotation of genes in terms of Gene Ontology categories and annotation of experimental evidence supporting the findings. This stage may also involve assignment of evidence codes to capture information about the source of experimental evidence.

The lessons learned from examining the database curation pipeline across a number of databases have particular importance for text mining and for the ability of biomedical researchers to extract information from the literature:

- Tools for managing the curation queue would be useful, but as curation criteria become more stringent (e.g., the article must have experimental evidence for a particular gene or protein), more human intervention is needed. Workflow and search technologies are needed here.
- Tools to keep curated data collections current would be useful. Such tools could provide an alert each time a new article is published that has information relevant to a

particular data collection. Even more useful would be the ability to flag new information that does not exist in the current information base. Agent based technologies could fill this need.

- Tools to locate relevant candidate passages within a full text article would be useful to curators, especially if these tools could be readily coupled to the ontology or terminology used for annotation. There is one such tool in use (Textpresso) that supports interactive curation and query, for specific model organism databases; there are also commercial tools coming into use for applications such as drug discovery. However, automated curation remains a difficult challenge as revealed by the BioCreAtIvE results, and more research is needed.
- The next generation of curation tools must support interactive curation: this is what the curators want and need; this is one of the stated goals for BioCreAtIvE 2.

8 BioCreAtIvE II

BioCreAtIvE is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. BioCreAtIvE arose out the needs of working biologists, biological curators and bioinformaticians to access the wealth of information in the literature, and to link this information to biological databases and ontologies. BioCreAtIvE focuses on the comparison of methods and community assessment of scientific progress, rather than on the purely competitive aspects. BioCreAtIvE is organized through collaborations between text mining groups, biological database curators and bioinformatics researchers.

The first BioCreAtIvE in 2004 addressed the detection of gene and protein names from text, their association to existing database entries and the extraction of protein annotations (i.e. protein - Gene Ontology concept associations). The Second BioCreAtIvE challenge (http://biocreative.sourceforge.net/index.html) will be held during October of 2006, with the workshop to be held in Spring 2007. It will consist of three tracks. The first will focus on finding the mentions of genes and proteins in sentences drawn from MEDLINE abstracts and is the same as Task 1A from BioCreAtIvE (Yeh et. al., 2005). The second track will involve producing a list of the EntrezGene identifiers for all the human genes/proteins mentioned in a collection of MEDLINE abstracts (Morgan et. al., 2007) and is similar to BioCreAtIvE I Task 1B (Hirschman, Colosimo et. al., 2005).

The third track of BioCreAtIvE II is a new advanced task on protein interaction detection, coordinated by a group under Alfonso Valencia at the CNIO (Spanish National Cancer Research Center), in collaboration with two of the main protein interaction databases (MINT and INTACT). The complexity of the first large scale proteomics experiments makes it an important application for text mining, to support the extraction of experimentally validated interactions from full text articles. A number of text mining tools are already accessible to the community, including the popular iHOP system developed by Valencia's group (www.ihop-net.org). The BioCreAtIvE protein interaction challenge will include detection of articles containing information relevant to protein interactions, the detection of actual protein interaction pairs, the extraction of experimental methods, and the corresponding text evidence. More than 55 teams are already training their systems for the protein interaction task, which will be evaluated using a test collection released in October 2006. We will then be in an excellent position to assess their performances and estimate the capacity of the current text mining systems.

The importance of the BioCreAtIvE methodology of evaluating text mining systems on real biological problems has attracted the interest of other groups developing curated databases. For example, we are supporting the OregAnno database developers in organizing a RegCreative annotation jamboree. The goal of this joint activity is to collect data to drive development and evaluation of interactive text mining tools applied to annotation of transcription factor binding sites (see http://www.dmbr.ugent.be/bioit/contents/regcreative/).

9 Lessons Learned and Impact

9.1 Lessons Learned

One lesson that is painfully apparent is the continued absence of any standards for organizing information on toxins and virulence factors. NIH/NIAID has funded eight Biodefense Research Centers, whose charter is to sequence the genomes of the CDC identified pathogens, but there is still no standard for collection of meta-data. There is also some work under the Genome Standards Consortium towards the capture of host-pathogen interaction information. The PAMGO (Plant Associated Microbe Gene Ontology) group has adapted the Gene Ontology to label certain kinds of host-pathogen interaction. However, our original goal of codifying virulence factors in terms of an appropriate ontology has not been met, because no such ontology exists. This is a glaring hole. Under the NSF funding for BioCreAtIvE, we plan to continue our interactions with key groups in this area.

A second set of findings come out of our experiments on interpretation of micro-array data. The rate of experimentation is increasing – it is now possible for researchers to generate more data than they can analyze through the use of high throughput techniques. This makes it even more critical to capture biological knowledge in computable form, e.g., in databases encoded using some kind of shared semantics or ontology. Our experiences in trying to apply biological data to interpret the experimental findings point to several key issues:

- Biological knowledge is sparse and data are highly distributed across many different databases, often in many different formats (and database-specific identifiers).
- Much time is spent simply translating from one database-specific form to another.
- As a result, workflow tools are essential to manage biological data.
- Capturing and organizing the meta-data and annotations associated with entities is critical, but very expertise-intensive and time consuming.
- Fortunately, the state of the practice is improving here; better resources are becoming available, and workflow management software is becoming available.

We have also gained more insight into the state-of-the-art for text mining, and we have made some substantial contributions to progress in this area. For existing text mining tools, there are a number of intrinsic, technical limitations:

- Entity tagging and identification. Entity tagging can be used effectively to index large collections, if a certain level of "noise" (misses and false alarms) can be tolerated (as in the drug discovery pipeline), or if the results can be manually curated. The accuracy of entity identification tools is improving and the tools work well for organisms with highly regular nomenclature (e.g., Worm, Yeast), but are not yet good enough to run in stand-alone mode in most cases, particularly for human genes or proteins.
- *Rapid Adaptation to New Tasks:* Each problem is distinct and requires special tailoring – which adds to the cost of an application. Text mining must overcome the

cost/performance barrier by creating modular tools that are easy to adapt to new requirements and new vocabularies through feedback mechanisms that support rapid tailoring and incremental learning. The Carafe tool developed by Wellner shows significant promise in providing a rapidly adaptable tool for entity normalization.

- *Curation Tools:* Better tools are needed to assist human experts in locating relevant information in articles and in mapping this information into the appropriate ontological classes or terminologies. An important goal is to use tools to speed up and improve the quality of manual curation. This will be a goal under the new NSF funded research.
- *Ontology Mapping and Maintenance:* Text mining tools are becoming useful in extracting information from free text and associating that information with the correct concepts in the ontology. Furthermore, text mining tools could support iterative improvements to ontologies by testing and highlighting new concepts as they are used for in applications. This will also be an area that we can investigate under the NSF funding, and possibly under funding from NCRR on biomedical ontologies.
- *Access to full text data*. Difficulty in accessing full text articles remains a stumbling block for indexing and text mining. As a result, indexing and search are often limited to PubMed abstracts. While this is starting to improve, it is still inhibits large-scale text mining activities.

Competitions and challenge evaluations will be an important means to address these intrinsic, technical limitations. They serve to bring together developers with the end users, e.g., biologists, annotators, and biomedical database developers. In particular, BioCreAtIvE, TREC Genomics, and the other challenge evaluations in the field have fostered the development and spread of tools for handling text data, including access to full text articles, comparative representation of results, and, most importantly, assessment of results by both automated means and by human experts. We are pleased to be international leaders in providing assessment of the state of the art for text mining applied to the biomedical area.

9.2 Impact

The major accomplishments under the MSR include the development of a pipeline for the analysis of high throughput micro-array data that has been applied to internal experiments.

In addition, under the MSR, we have established MITRE as a major player in bioinformatics, particularly in the area of text mining for biomedical literature, but with increasing visibility in the area of curated biological databases. In the course of the MSR, the staff has published 20 articles (see Appendix A).

We have played a major role in organizing the biomedical text mining community, including organizing or chairing ten events over the three years. In particular, we not only founded BioCreAtIvE, in collaboration with Prof. Alfonso Valencia, now at the Spanish National Cancer Research Center, but we also co-founded BioLINK, the SIG for text mining in Biology.

We have also participated in a number of advisory boards and standards bodies, including serving on the Gene Ontology Advisory Board (Hirschman); participating in the IEEE Bioinformatics Standard (Morgan); serving on the National Center for Biomedical Ontologies Evaluation Committee (Mani); serving on the TREC Genomics Advisory Board (Colosimo, Morgan, Hirschman); and serving on the NIH-funded Maine IdEA Network for Biological Research Excellence Advisory Board (Hirschman). In addition, Hirschman and Mani have both served on NIH funding panels, and Hirschman and Colombe have served on NSF funding panels for bioinformatics.

Appendix A Publications

1. Lynette Hirschman, Alex Morgan, Alexander Yeh. "Rutabaga By Any Other Name: Extracting Biological Names," The Journal of Biomedical Informatics: 2002 Aug; 35(4):247-59. (Best Paper Award, 2003).
2. Marc E. Colosimo, Susan Tran, and Piali Sengupta. "The Divergent Orphan Nuclear Receptor ODR-7 Regulates Olfactory Neuron Gene Expression via Multiple Mechanisms in Caenorhabditis Elegans," (2003) Genetics 165(4):1779-91.
3. Colosimo M.E., Brown A., Mukhopadhyay S., Gabel C., Lanjuin A.E., Samuel A.D., Sengupta P.: "Identification of Thermosensory and Olfactory Neuron-specific Genes via Expression Profiling of Single Neuron Types." Cur Biol. 2004 Dec 29;14(24):2245-51.
4. Murray Wolinsky, Jian Song, Jason Gans, Cathy Cleland, Robert Leach, Chris Stubben, Yan Xu, Luther Lindler, Kevin Anderson, Elliot Lefkovitz, Alexander Morgan, Marc Colosimo, Alexander Yeh, and Lynette Hirschman "TVFacDB: A Comprehensive Microbial Toxins and Virulence Factors Database," Proceedings of BTR 2204: Unified Science and Technology for Reducing Biological Threats and Countering Terrorism, Albuquerque, NM: March, 2004.
5. Alexander A. Morgan, Lynette Hirschman, Marc Colosimo, Alexander Yeh, Jeff Colombe. "Gene Name Identification and Normalization Using a Model Organism Database," The Journal of Biomedical Informatics, 2004 Dec;37(6):396-410.
6. L. Hirschman, C., Blaschke, A. Valencia, "The BioLink SIG Workshop at ISMB 2004," Comp Funct Genom 2005; 6:58-60.
7. Kim K., Colosimo M.E., Yeung H., Sengupta P., "The UNC-3 Olf/EBF Protein Represses Alternate Neuronal Programs to Specify Chemosensory Neuron Identity," Dev Biol. 2005 Oct 1;286(1):136-48.
8. Colosimo, M., Hirschman, L., Morgan, A., Yeh, A., "Data Preparation and Interannotator Agreement BioCreAtIvE Task 1B," BMC Bioinformatics 2005, 6(Suppl 1):S12.
9. A. S. Yeh, Lynette Hirschman, Alexander A. Morgan, Marc Colosimo. "BioCreAtIvE Task 1A: Gene Mention Finding Evaluation," BMC Bioinformatics 2005, 6(Suppl 1):S2 (24 May 2005).
10. Lynette Hirschman, Marc Colosimo, Alexander A. Morgan, Alexander S. Yeh. "Overview of BioCreAtIvE task 1B: Normalized Gene Lists," BMC Bioinformatics 2005, 6(Suppl 1):S11 (24 May 2005).
11. B. Wellner, "Weakly Supervised Learning Methods for Improving the Quality of Gene Name Normalization Data," Proc of ACL BioLINK Workshop, June 24, 2005.

12. B. Wellner, J. Castaño, J. Pustejovsky, "Adaptive String Similarity Metrics for Biomedical Reference Resolution," Proc of ACL BioLINK Workshop, June 24, 2005.
13. Blaschke, C., Yeh, A., Camon, E., Colosimo, M., Apweiler, R., Hirschman, L. and Alfonso Valencia, "Do You Do Text?" Bioinformatics 2005 21(23):4199-4200 (2005).
14. Hirschman, L., Damianos, L. "Mining Online Media for Global Disease Outbreak Monitoring," in AMS-DIMACS volume on Epidemiology, Eds. James Abello and Graham Cormode, American Mathematical Society (AMS), 2006.
15. Cohen, K.B., Bodenreider,O. and L. Hirschman, "Linking Biomedical Information Through Text Mining: Session Introduction," Pacific Symposium on Biocomputing 11:1-3(2006).
16. L. Hirschman and C. Blaschke, "Evaluation of Text Mining in Biology, Text Mining in Biology and Biomedicine," eds S Ananiadou and J McNaught, 2006. (book chapter)
17. Hirschman, L., Hayes, W.S., and A Valencia, "Knowledge Acquisition from the Biomedical Literature," to appear in Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences," Christopher J. O. Baker and Kei-Hoi Cheung, eds. (book chapter)
18. Leo Obrst, Werner Ceusters, Inderjeet Mani, Steve Ray, and Barry Smith. Evaluation of Ontologies: Toward Improved Semantic Interoperability. To appear in Christopher J. O. Baker and Kei-Hoi Cheung, Eds., Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Springer. (book chapter)
19. Morgan, A., Wellner, B., Colombe, J., Arens, R., Colosimo, M., and Hirschman, L. "Evaluating Human Gene And Protein Mention Normalization To Unique Identifiers," accepted by Pacific Symposium on Biocomputing, 2007.
20. Wellner, B. Huyck, J., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L, Yeh, A. Hitzeman J., and L. Hirschman, "Rapidly Retargetable Approaches to De-identification," to appear in the American Medical Informatics Association Workshop on Natural Language Processing, Baltimore, November 2006.

Appendix B Presentations and Tutorials

Briefing to Ron Walters, ITIC April 2004.

Briefing to Dale Nordenberg, CIO of CDC, Sept 2003.

M. Colosimo, "Critical Assessment of Information Extraction Systems in Biology (BioCreAtIvE)" Biocurators Meeting 2003, Milwaukee, WI.

A. Yeh, L. Hirschman, A. Morgan, M. Colosimo, "BioCreAtIvE Task 1A: Gene-Related Name Mention Finding Evaluation," Invited talk, BioCreAtIvE Workshop, Granada, March 2004.

L. Hirschman, M. Colosimo, J. Colombe, A. Yeh, A. Morgan (Invited Talk) "Normalized Gene List Extraction," BioCreAtIvE Workshop, Granada, March 28-31, 2004.

M. Colosimo, L. Hirschman, A. Morgan, A. Yeh, J. Colombe, "Data Preparation and Interannotator Agreement, BioCreAtIvE Task 1B," BioCreAtIvE Workshop, Granada, March 28-31, 2004.

Marc Colosimo, "Report on the 2004 BioCreAtIvE Workshop," BioCurator's Meeting, Eugene Oregon, September 2004.

Lynette Hirschman (Invited Talk), "Naming, Describing, Classifying – Ontologies for Biological Databases," Bioinformatics 2004, Linköping, Sweden, June 3-6, 2004.

Lynette Hirschman (Invited Talk), "Extracting Computable Semantics: Text Mining and Ontologies for BiologyE-Biosci/Ariel Meeting," Hinxton, UK, October 2004.

A. Morgan (Invited Talk), "Linking Text Mentions to Biological Identifiers," Ontario Centre for Genomic Computing, Text Mining Tools for Bioinformaticians and Biologists Workshop, February 4, 2005.

A. Morgan (Invited Talk), "Linking Text Mentions to Biological Identifiers," Japan's National Institute for Informatics and the University of Tokyo, e-Biology Initiative: Towards New Frontiers of Biology, March 11, 2005.

L.Hirschman (Invited Talk), "Portability and Domain Models: Biology as a Case Study for Information Extraction," DHS Text Analysis Workshop, May 2005.

Lynette Hirschman (Invited Talk), "Mapping from Text to Ontology for Biological Applications, Knowledge-Based Bioinformatics Workshop," Montreal, September 21-23, 2005.

Lynette Hirschman (Keynote), "Evaluating What Biologists Want," Symposium on Semantic Mining in Biomedicine," Jena, April 10-12, 2006.

Alex Morgan, "2nd Annual International Symposium on Semantic Mining in Biomedicine," Jena, Germany, 2006: Invited Tutorial on the Evaluation of Text Mining Systems (with Martin Krallinger).