From the Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

Applied Bioinformatics for Gene Characterization

Raf M. Podowski

Stockholm 2006

Podowski, Raf M.
Applied bioinformatics for gene characterization

ISBN 91-7140-818-5
Printed by Akademitryck AB, Valdemarsvik 2006 www.akademitryck.se
Graphic design by Akademitjanst, www.akademitjanst.se

Abstract

As the vast majority of human protein-encoding genes are now identified, the new challenge placed before life scientists is the determination of the functions of the proteins. Traditionally, intense, directed efforts are applied to decipher the function of a novel protein using laboratory techniques. Currently, increasing efforts are directed at the generation of high-throughput results for large numbers of genes using new technologies and the application of robotics to established methods. These efforts can generate large, complex, often noisy datasets, which are difficult to interpret. The extraction of information from genomics data that is relevant for specific scientific research efforts is required to accelerate functional characterization and annotation of genes by the scientific community.

The research presented in this thesis highlights and addresses deficiencies in gene/protein function annotation. The bioinformatics tools and methodologies presented share the common theme of facilitating research scientists with means to understand and to interpret gene-specific data. The work, which addresses both diverse types of genomics data and a broad set of computational approaches, is united by the hypothesis that computational approaches to genomics data analysis can assist in the characterization of human protein-encoding genes.

The initial sections of the thesis describe the identification of human protein-encoding genes for which there is little or no functional annotation. The initial chapter presents the first bioinformatics method for quantifying the level of annotation of individual genes and monitoring progress. We developed the first Gene Characterization Index, a computational method for scoring the extent to which each human protein-encoding gene is functionally described. Inherently a reflection of human perception in a window of time, the Gene Characterization Index serves both as a tool for assessing the novelty of individual genes, and for the assessment of short-term annotation progress on a genome scale. Based on the surveyed opinions of life scientists, machine learning methods are used to identify annotation properties which correlate with the expressed opinions. The characterization score enables researchers to highlight poorly characterized genes for which focused efforts can be made to extract information from genomics data. The procedure was subsequently applied to study the temporal changes in gene characterization

over recent years, both to identify poorly and well characterized genes within pharmaceutically relevant protein classes, and to highlight poorly characterized genes for which gene-specific patent applications exist, with potentially useful annotations.

In the second paper, computational approaches are used to identify specific protein families that share evolutionarily-conserved domains for which the biochemical function is unknown. For the identified domains, a gene-centric data centre, NovelFam3000, is created to facilitate shared annotation of protein function. This software system allows for communal annotation, both experimental and computational, of individual proteins. Once a domain is characterized in one protein, the presence of a similar sequence in an uncharacterized protein serves as a basis for inference of function. Thus, knowledge of a domain's function, or the protein within which it arises, can facilitate the analysis of the entire set of proteins.

The subsequent sections of the thesis focus on the creation of bioinformatics methods to assist human interpretation of gene function. Interpretation of large-scale biological data can be aided by visualization—humans can perform complex interpretation of data through visual assessment. Heatmaps have emerged as a preferred technique for the display of genomics data, as they provide an extra dimension of information in a two dimensional display. However, an increasing focus on the integration of data from multiple sources has created a need for the display of additional dimensions. In order to improve the identification of relationships between co-expressed genes identified in microarray-based experiments, the Parallel HeatMap viewer was developed for four-dimensional data display. The flexible data entry structure of the Parallel HeatMap viewer facilitates the display of both continuous and discrete data. The Parallel HeatMap viewer enables knowledgeable life science researchers to observe patterns and properties within high-throughput genomics data in order to rapidly identify biologically logical relationships.

Researchers seeking to understand gene function often turn initially to the scientific literature. Hindered by a historic lack of standard gene and protein-naming conventions, they endure long, sometimes fruitless literature searches. The final chapter of the thesis focuses on the computational identification of abstracts which may be relevant to gene function—the essential and difficult challenge that must be overcome for computational assisted literature review. A gene symbol, such as CAT, can refer to any one of a number of distinct genes, as well as to numerous non-gene entities, and its correct sense cannot be easily distinguished in text without a close examination. The final chapter introduces a computational approach, SureGene, to aid in addressing this "disambiguation problem". The system is based on supervised machine learning, resulting in a distinct model for the identification of rel-

evant abstracts for each gene. The training sets for all genes are extracted automatically from functional descriptions and MEDLINE references in the Entrez Gene and SwissProt databases. The system was able to achieve high quality gene disambiguation using scalable automated techniques.

This thesis explores the hypothesis that computational methods can facilitate the identification and characterization of poorly annotated genes. The bioinformatics approaches to this problem assist researchers in advancing our understanding of the functional of human protein encoding genes.

Original Publications

I

Podowski, R.M., Kemmer, D., Brumm, J., Wahlestedt, C., Lenhard, B., Wasserman, W.W.
Gene Characterization Index: A Metric for Accessing How Well We Understand Our Genes.
To be submitted to Nature Biotechnology (2006)

II

Kemmer, D., **Podowski, R.M.**, Lim, J., Arenillas, D., Hodges, E., Roth, P., Sonnhammer, E.L.L, Höög, C., Wasserman, W.W.
NovelFam3000 – Uncharacterized Human Protein Domains Conserved Across Model Organisms.
BMC Genomics **7**:48 (2006)

III

Podowski, R.M., Miller, B., Wasserman, W.W.
Visualization of Complementary Systems Biology Data with Parallel Heatmaps.
IBM Journal of Research and Development **50**:6 (2006)

IV

Podowski, R.M., Cleary, J.G., Goncharoff, N.T., Amoutzias, G., Hayes, W.S.
SureGene, a Scalable System for Automated Term Disambiguation of Gene and Protein Names.
Journal of Bioinformatics and Computational Biology **3**(3):743–770 (2005)

Related Publications

Kemmer, D., Faxen, M., Hodges, E., Lim, J., Herzog, E., Ljungström, E., Lundmark, A., Olsen, M.K., **Podowski, R.M.**, Sonnhammer, E.L.L, Nilsson, P., Reimers, M., Lenhard, B., Roberds, S.L., Wahlestedt, C., Christer Höög, C., Pankaj Agarwal and Wasserman, W.W.
Exploring the Foundation of Genomics: A Northern Blot Reference Set for the Comparative Analysis of Transcript Profiling Technologies.
Comparative and Functional Genomics **5**(8):584–595 (2005)

Kutsenko, A.S., Gizatullin, R.Z., Al-Amin, A.N., Wang, F., Kvasha, S.M., **Podowski, R.M.**, Matushkin, Y.G., Gyanchandani, A., Muravenko, O.V., Levitsky, V.G., Kolchanov, N.A., Protopopov, A.I., Kashuba, V.I., Kisselev, L.L., Wasserman, W.W., Wahlestedt, C., Zabarovsky, E.R.
NotI flanking sequences: a tool for gene discovery and verification of the human genome.
Nucleic Acids Research **30**(14):3163–3170 (2002)

Zabarovska, V.I., Gizatullin, R.Z., Al-Amin, A.N., **Podowski, R.M.**, Protopopov, A.I., Lofdahl, S., Wahlestedt, C., Winberg, G., Kashuba, V.I., Ernberg, I., Zabarovsky, E.R.
A new approach to genome mapping and sequencing: slalom libraries.
Nucleic Acids Research **30**(2):E6 (2002)

Podowski, R.M., Sonnhammer, E.L.L.
MEDUSA: Large Scale Automatic Selection and Visual Assessment of PCR Primer Pairs.
Bioinformatics **17**(7):656–7 (2001)

Kashuba, V.I., Protopopov, A.I., **Podowski, R.M.**, Gizatullin, R.Z., Li J., Klein, G., Wahlestedt, C. and Zabarovsky, E.R.
Isolation and chromosomal localization of a new human retinoblastoma binding protein 2 homologue 1a RBBP2H1A.
Eur. J. Human Genetics **8**(6):407–413 (2000)

Protopopov, A., Kashuba, V., **Podowski, R.M.**, Gizatullin, R., Sonnhammer, E, Wahlestedt, C. and Zabarovsky, E.R.
Assignment of the GPR14 gene coding for the G-protein-coupled receptor 14 to human chromosome band 17q25.3 by fluorescent in situ hybridization.
Cytogenetics and Cell Genetics **88**(3–4):312–313 (2000)

Zabarovsky, E.R., Gizatullin, R.Z., **Podowski, R.M.**, Zabarovska, V., Xiu, L., Muravenko, O., Kozyrev, S., Petrenko, L., Skobeleva, N., Li, J., Protopopov, A., Kashuba, V., Ernberg, I., Winberg, G. and Wahlestedt, C.
NotI clones in the analysis of the human genome.
Nucleic Acids Research **28**(7):1635–1639 (2000)

Contents

List of abbreviations

ACM	Association for Computing Machinery
DNA	Deoxyribonucleic Acid
EST	Expressed Sequence Tag
GCI	Gene Characterization Index
GDS	GEO Data Sets
GEO	Gene Ontology Omnibus
GO	Gene Ontology
HMM	Hidden Markov Model
HUGO	The Human Genome Organization
IEEE	Institute of Electrical and Electronics Engineers, Inc.
IR	Information Retrieval
KEGG	Kyoto Encyclopedia of Genes and Genomes
MARS	Multivariate Additive Regression Splines
MeSH	Medical Subject Headings
MIPS	Munich Information Center for Protein Sequences
MOA	Method of Action
NLP	Natural Language Processing
PHM	Parallel HeatMap
RNA	Ribonucleic Acid
SVM	Support Vector Machines
TF	Transcription Factor
TFBS	Transcription Factor Binding Site

Preamble

Most of what you get taught is lies. It has to be. Sometimes if you get the truth all at once, you can't understand it.
Terry Pratchett

Human gene annotation

The number of protein-encoding human genes identified has reached a plateau [1], leaving researchers with the challenging task of ascribing biochemical function(s) for each protein [2]. Broad genome sequencing and functional genomics studies, partially motivated by the goal to discover the functions of uncharacterized proteins, have provided a distributed set of data collections suitable to catalyze the inference of the functions of proteins.

In contrast to Mendelian views of a gene as a trait carrier, a bioinformatics perspective views a gene entity as a combination of the DNA, RNA and protein product, with physical and biochemical properties that can be studied for functional characterization.

Gene characterization approaches

Elucidation of the function(s) for each human protein-encoding gene is a prominent challenge in biomedical research. Systematic characterization projects are underway, ranging from the ENCODE project for detailed genome annotation [3] to the phenome projects to identify phenotypes generated by mutations of human gene orthologs in model organisms [4–6]. These efforts are undertaken, in part, to evoke new insights into the functions of uncharacterized genes revealed through the successful sequencing of the human genome. At the level of basic human curiosity, scientists are drawn to these uncharacterized genes, for it is the deciphering of the functions of these genes which offers the greatest potential to gain fundamental insights into biological processes; to peer into the unknown. The therapeutic and financial benefits associated with successful identifi-

cation of genes that are suitable targets for pharmaceutical research and informative biomarkers for treatment selection stands as another strong motivator.

The arsenal of the modern molecular researcher, when directed at specific genes, can quickly elucidate properties that offer glimpses of underlying functions. In the laboratory we can determine where the encoded protein localizes within the cell, the spatio-temporal coordinates of gene activity, the function in cells or model organisms through biological assays, and further techniques ad infinitum. To unleash these expensive and time consuming studies, researchers (and funding agencies) are often motivated by preliminary glimmers of functional knowledge. As established by the wisdom of the international yeast research community to systematically knock-out each yeast gene in turn [7], and the realization that the number of uncharacterized human genes is dramatically less than anticipated [8], a very different paradigm has emerged. Rather than focusing resources upon genes illuminated by the sparks of preliminary research results, those genes that have remained shrouded in darkness may be brought forward for study. In the ENCODE project, undertaken by a portion of the global research community to systematically annotate functions for 1% of the human genome, a portion of the genome was selected for study for the glaring absence of knowledge about the genes in the region. The Allan Brain Atlas [9] places a premium on the systematic study of expression in the mouse brain of uncharacterized genes. In the biotechnology and pharmaceutical industries, gaining insights into the functions of uncharacterized genes can offer a direct and meaningful path to successful patent applications. In short, it is now acceptable to focus attention and resources on the diminishing set of uncharacterized human genes.

Information about individual genes is available from a variety of sources. A number of high-profile data centers such as Entrez Gene [10], GeneCards [11] and SwissProt [12] provide curated, comprehensive information with links to individual information sources. A researcher can search through these resources for genes associated with a disease or phenotype. Discovery or prediction of functions of uncharacterized genes is not directly possible. In order to predict the roles of proteins with some confidence, one can study interactions or

associations with other genes. For example, Aerts et al. [13] introduced a ranked gene prioritization method which integrates multiple genomic data sources. The Endeavour system creates a model representing the most predominant characteristics for a set of training genes, selected by the user as positive examples of a function or process of interest. Multiple resources are queried to derive disease or pathway information for the training genes including literature, functional annotation, microarray and EST expression, protein domains, protein-protein interactions, pathway membership, *cis*-regulatory modules, transcriptional motifs and sequence similarity. Additional resources can be added by the user. The model is then applied to the entire database of genes to select those most similar to the genes in the training set. By merging heterogeneous data sources through rigorous statistical methods, an overall ranking of test genes is generated relating them to the disease or biological process of interest. A number of studies have demonstrated that consistent results for interactions between homologous genes in multiple organisms, so called Interolog Analysis, can be lead to discovery of reliable characterization information not otherwise possible [14–16]. Therefore, human protein characterization efforts that focus on similar proteins across multiple organisms are expected to more effectively capitalize on the available genomics data. This powerful approach has one limitation - sufficient information about each gene must exist to allow for the inference of gene-gene association.

Genomic data visualization tools

Systems biology research generates large, complex datasets. While clustering algorithms can identify subsets of genes that behave similarly, interpretation of inter-gene relationships can be difficult. In only a small subset of cases can a biological theme be accurately ascribed to a statistical grouping of genes. Interpretation is complicated by the fact that popular clustering algorithms such as hierarchical, K-means and self-organizing maps [17], are guaranteed to produce clusters, even if no underlying biological process or statistical motivation exists. Assessment and interpretation of clusters can be sim-

plified when data from multiple sources is correlated. For example, Robinson, et al. [18], combined gene expression data for 237 distinct gene knockouts [19] with information from the Munich Information Center for Protein Sequences (MIPS) Functional Classification database [20] to identify a 76-gene cluster involved in amino acid biosynthesis and metabolism. A number of bioinformatics tools that facilitate such integrative approaches exist. INCLUSive [21] is a system analyzing gene clusters from expression microarray data for transcription factor (TF) binding motif identification. GoMiner [22] utilizes expression levels and Gene Ontology (GO) [23] for organization and biological interpretation of genes with respect to a selected subset of user selected genes. EASE [24] identifies over-represented attributes for a gene list or cluster using multiple gene annotation data sources such as GO and MeSH [25] terminology, transcription factor binding sites, protein domains, pathways and chromosomal locations. FunSpec [18] identifies common attributes for a given gene set based on information from sources such as GO, Pfam [26], MIPS and a number of specific high-throughput data sources. The oPOSSUM system [27] searches for evidence of co-regulation by one or more transcription factors, combining a pre-computed database of conserved transcription factor binding sites in human and mouse promoters with statistical methods for identification of sites over-represented in a set of co-expressed genes.

Methods based on comparative analysis across species (co-expression networks and interologs) are becoming common. Interolog Analysis is based on observation of mutually consistent interactions in multiple species (Figure 1). Such methods require a high quality map enumerating gene homology relationships among species. Matthews et al. [28] investigated the extent to which a protein interaction map generated in one species can be used to predict interactions in another species by using *S. cerevisiae* two-hybrid interaction maps to predict interactions in *C. elegans*. A complementary study by van Noort, et al. [29], investigated gene function prediction by conserved co-expression in yeast and worm orthologs. In a broader study, Stuart, et al. [30], constructed a gene co-expression network based on gene expression data and ortholog sequence similarity among model organisms of human, worm, fly and yeast. Uniting the Atlas

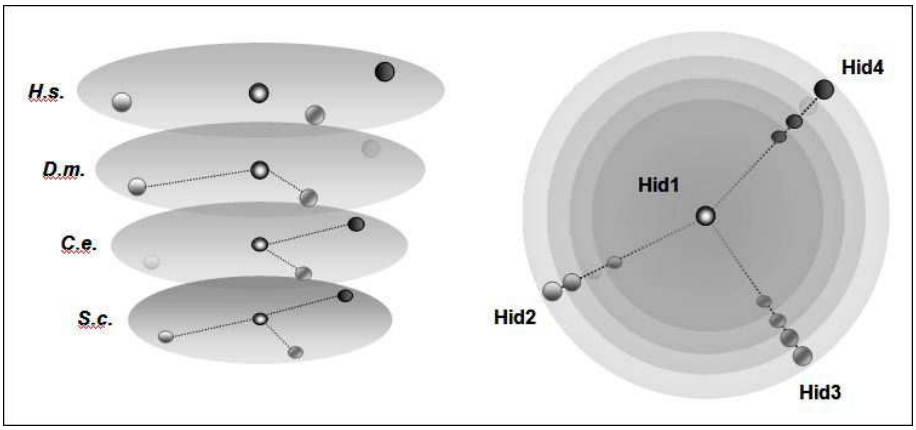

Figure 1. Interolog Analysis. Interolog mapping of conserved protein networks across four model species: human (H.s.), fly (D.m.), worm (C.e.) and yeast (S.c.). Individual species protein interaction networks are overlaid and aligned based on orthologous genes designated by HomoloGene Ids (Hid), producing an extrapolation of the complete interaction network. Confidence in the network grows as more evidence is discovered supporting pair-wise interactions in each species. In the above example, the interaction between Hid1 and Hid3 is supported in 3 species, while those of Hid1-Hid2 and Hid1-Hid4 have support in just 2 species each.

[31] database and HomoloGene [32] with interactive visualization, the Ulysses system [33] performs interolog analysis for the parallel analysis and display of protein interactions detected in model organisms, including human, worm, fly and yeast.

Visualization tools for assessment of correlations offer an alternative approach based on accessing the cumulative knowledge of human specialists – knowledge that can be difficult to replicate computationally [34]. Computer-assisted data visualization has been an active area of research since the early 1960's. With the influx of high-throughput data, the scientific community has recognized the benefits of multi-dimensional visualization in data exploration and interpretation. Numerous publications, workshops and conferences addressing these issues can be found through the IEEE Computer Society [35] and ACM [36] web sites. Improved coordination across multiple views has received recent emphasis [37]. Boukhelifa and Rodgers [38] describe a model for expressing coordination in multiple view visualization systems. Ross, et al. [39] created a set of tools for profiling the performance of individual visualization components.

Bertini, et al., [40] developed VidaMine, a data exploration and mining system exploiting multiple views. Researchers at the University of British Columbia created graphical tools for comparison of phylogenetic trees and multiple sequence alignments, utilizing progressive rendering and localized zooming, including TreeJuxtaposer [41] and SequenceJuxtaposer [42]. These tools concentrate on assessment of only a single data source albeit, in multiple views.

There exist tools specifically designed to explore a single source of biological information, as exemplified by Osprey [43] for navigation of molecular interaction data, Pathway Voyager [44] for KEGG [45] metabolic pathway visualization, and the Reactome knowledgebase of biochemical pathways and biological processes browser [46]. Cytoscape, a continually evolving network visualization program [47], generates an interactive graph of molecular interaction data with the ability to assign attributes such as Gene Ontology definitions and gene expression level information to the nodes representing individual molecules. Likewise, graph edges can be assigned labels or numerical values representing pair-wise interaction type or strength.

As the availability of complementary high-throughput data is growing, the means to visually discover new relationships within large and complex data has become critical. Heatmaps are well established in genomics, provide a means to rapidly identify relationships across large datasets, and conveniently display continuous data through color intensity (Figure 2). In 1997, Weinstein, et al. [48], explored the relationship between 3989 compounds and 76 molecular targets believed to interact in cancerous cell lines, by correlating gene expression of thriving tumor cells and the levels of growth inhibition brought about by a potentially therapeutic compound. The results

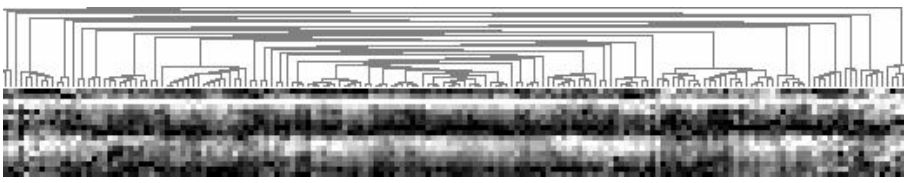

Figure 2. Heatmap. Microarray gene expression results have been clustered hierarchically showing distinct patterns of over-expression (bright) and under-expression (dark). Each row represents an individual experiment and each column an individual gene.

were presented as a heatmap, clearly distinguishing which compounds were active in cell lines with elevated expression levels of the target proteins. In 1998, Eisen, et al. [49], presented a system for the analysis of genome-scale, multi-experiment, microarray expression data. In one example, genes were clustered hierarchically based on the expression patterns for growth response in human cells, gathered

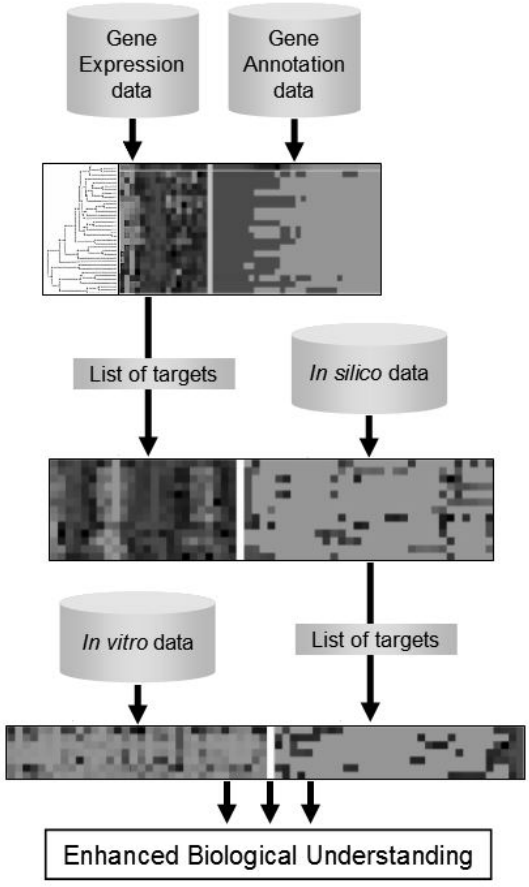

Figure 3. Multi-Dimensional Data Visualization. Benefits of multi-dimensional data visualization can lead to new insight into the biological process of a gene. Gene clusters from microarray expression experiments can be characterized based on existing gene annotation resources. A selected list of target genes is assessed against in silico information such as transcription factor binding predictions. Finally, predictions for selected genes are compared against in vitro data from protein-protein interaction experiments, strengthening confidence in functional aspects of individual genes.

from 12 samples over a 24 hour period of time. Functionally related clusters were identified including those of genes involved in cholesterol biosynthesis, cell cycle, immediate-early response, signaling and angiogenesis, and wound healing. They concluded that genes of similar function cluster together. Subsequently the majority of studies of genome-sized expression data have utilized heatmaps and/or line plots, as exemplified by Gasch [50] and Wen NAME [51].

Most gene expression analysis packages, from academic tools such as Cluster and TreeView [49], Hierarchical Clustering Explorer (HCE) [52, 53], GDS Browser [54] and Prism [55], to commercial products such as ArraySCOUT [56], GeneSpring [57] and Spotfire DecisionSite for Functional Genomics [58], provide heatmap visualization tools that are linked to clustering algorithms [59]. Layering of complementary data into the visual display, however, has been limited. The above-mentioned tools provide visualization and annotation enhancements to support cluster analysis, including dendrograms, scatterplots, line graphs, detailed row and column descriptions and links to external annotations. Similarities between Gene Ontology annotations assigned to individual genes [52, 60, 61, 62] can provide a useful, albeit limited, hint at inter-gene relationships. While multidimensional visualization tools for database exploration are long established [63, 64], there are no established bioinformatics tools that facilitate the visualization of functional relationships from unrelated sources on a global scale as envisioned in Figure 3.

Literature searches for gene characterization information

Heatmaps and other visualization approaches are based on assessing the specific domain knowledge of a scientist. Ultimately much of this knowledge is drawn from the primary scientific literature. In order to gain large-scale access to data in published papers, text mining methods must be created to link functions/attributes with genes. To initiate a text analysis procedure for the discovery of gene-gene associations, the first step is to select documents which address a given gene.

Text mining is a systematic process involving multiple steps (Fig-

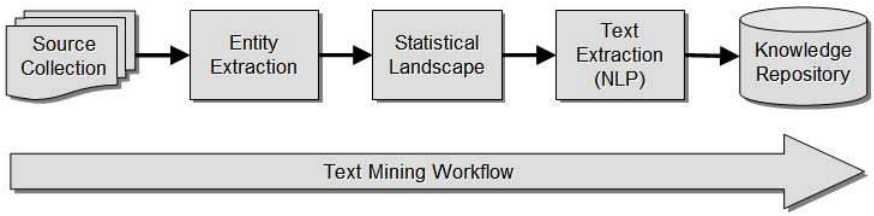

Figure 4. Text Mining Workflow. Typical text mining workflow involves identification and subsequent collection of document sources, biological/biochemical/medical entity extraction, statistical content analysis, including co-occurrence statistics, document clustering and classification, and natural language processing enhances by domain-specific ontologies. Comprehensive results are stored in databases for user queries and computational analysis.

ure 4). Starting with identification and access to a collection of sources, biological, biochemical or medical entity identification and extraction form the text follows. Statistical content analysis provides a landscape of information useful for over-association studies, document, concept or entity clustering and machine learning classifiers for automatic tagging of documents with human-generated themes or topics. Natural Language Processing, with the aid of targeted ontologies is used to derive directed, factual relationships between entities. Ontologies are rich descriptions of the concepts and relationships that exist within a domain. They can be used to generate controlled vocabularies, thesauri and taxonomies. This information is stored in a database repository and utilized for a number of automated processes such as pathway or protein interaction network generation.

Finding relevant information

Text searching represents the first level of text mining. It is more formally known as information retrieval (IR). The resulting output is usually presented as ordered document lists, with documents ranked according to a keyword-based scoring function. Because documents are treated merely as "bags of words," all context and semantic variation is ignored. Therefore, keyword searches tend to return a high volume of "hits" with little ability to discriminate nuance, complex connections, or even the relevance of the concepts communicated in the document in which the keyword resides. On the other hand, key-

word-based text search is the most popular form of text mining because it is the most familiar. A researcher performs a search and then analyzes the results manually. One can also run multiple searches looking for intersections between literature sets using logical functions [65].

The first major drawback of keyword searching is that the task of sorting the search engine output falls on the investigator, which exploits neither the computer's nor the human's strengths. In addition, keyword searches suffer from polysemy (the same word having different meaning in different contexts), which requires the reader to examine documents for relevance, where a large number may be completely incorrect, and synonymy (multiple words referring to the same concept), which requires the investigator to know (and employ) all possible alternative synonyms to ensure a complete search. Synonymy is a particularly difficult problem in biology literature, where proteins routinely have many names and abbreviations often overlap common English words.

Assuming that the gene-document mapping problem can be overcome, numerous approaches to gene characterization become feasible. Three common approaches are (i) statistical over-representation of terms in a set of documents for a gene or gene-set; (ii) text categorization of documents to determine if a category is over-represented in the set of gene or gene-set related documents; and (iii) document clustering to facilitate organized review of documents for a gene or gene-set.

Functional analysis of a gene set with a possible preexisting relationship, such as a co-expressed, microarray-derived gene cluster, can be facilitated by statistical text mining. A literature set associated with a gene cluster can be analyzed for statistically significant frequency differences for annotation terms from vocabularies such as GO and MeSH, or entity classes such as genes, diseases, biological processes, or pathological processes. Gene co-occurring term analysis has been integrated with a gene-expression clustering algorithm itself to assist in the determination of gene clusters [66]. Another example is TXTGate [67], a system using statistical analysis of MEDLINE [68] literature with biomedical domain vocabularies for the analysis of a group of genes. It generates a gene cluster profile consisting of statis-

tically over-represented terms and phrases associated with the genes under consideration.

Text categorization is a content classification method, requiring user curation of a model. By defining the properties of a topic of interest, the user enables a computational identification of documents addressing this topic. Such an approach requires a notion of relevant attributes to study. Automatic text categorization for text mining is usually employed to assign documents to specified subsets. Text categorization is a supervised machine-learning technique, requiring training data to generate the category models. It can be used to generate large taxonomic structures of documents similar to Yahoo's [69] classification of Web pages. On the whole, text categorization needs to be highly focused and customized to the project. Text categorization, more than any other area of text mining, requires very careful thought and design. Extensive testing is required to understand the behavior of the categorization model.

Document clustering is used for knowledge discovery and provides a hint of the diversity of themes within an otherwise uncharacterized document collection. In particular, clustering is used when exploratory searches result in hundreds or thousands of documents. Clustering can greatly improve the grouping and prioritization of a set of documents. In the context of text analysis, document clustering arose from a desire to improve information retrieval systems [70, 71], identify similar documents [72], and better organize and browse a group of documents [73, 74]. Document clustering is a form of unsupervised machine learning [75]. At its simplest and purest level, document clustering requires no prior knowledge or expectations about the contents and provides concept extraction [76] and knowledge navigation. Furthermore, concept extraction can seed an automatic derivation of classifications and serve as a method of "ontology induction" [77]. Traditional clustering methods rely less on semantic analysis and instead utilize multivariate statistical techniques to form clusters of similar objects in a multidimensional space [78]. In every case, the process involves generation of characteristic document vectors. Most frequently, these vectors are based on individual word frequencies in the document. To reduce the significance of frequently occurring words found in a majority of the documents, such as com-

mon English words, normalization or weight schemes can be applied: inverse document frequency, probabilistic weights, stoplists (a list of specific words that will be excluded from the analysis), or domain-specific weighted theme lists. At this stage, a method may diverge from purely automatic clustering toward semi-supervised, partially categorization-based cluster identification. Document vectors are used for calculating a similarity (or distance) metric between two documents or a document and a cluster centroid (a vector representing the center of a cluster of documents).

As introduced above, the capacity to perform text-based analysis is fundamentally linked to the gene disambiguation problem – we must be able to identify which genes are addressed in a sentence or document.

Gene naming convention

Prior to the establishment of high-throughput gene sequencing methodologies, new genes were discovered infrequently, after great effort. Naming the newly discovered gene was undertaken by the discovering team with little thought of the impact of the name for the future. Fruit fly researchers are best known for extreme naming practices, with names such as 'lot', 'sarah', 'ken and barbie', 'lost in space', 'cheap date', 'drop dead' and 'swiss cheese'. Frequently, genes of unrelated function in even a single species were given identical names by their respective discoverers. This did not pose a large problem while isolated communities of researchers devoted themselves to studying a small subset of genes within a single organism. With a growing number of cross-species studies and increasing importance of intra-species network interaction analysis, haphazard naming conventions have created a significant problem for automated analysis methodologies, especially those relying on literature analysis of unstructured text.

In 1979, the HUGO Gene Nomenclature Committee (HGNC) [79] became an official body for approval and implementation of human gene names and symbols. Initial guidelines for nomenclature of human genes were published, followed by a number of updates, the most recent in 1997. All approved human gene symbols are ac-

cessible from the HGNC database. As of June 2006, the database contains 23,422 official symbols for genes. Over the history of the project, 4457 entries have been changed at least once. At present, 14943 genes contain one or more aliases. 78 former symbols and 358 aliases are currently official gene symbols for other genes. Overall, HUGO maintains 53906 unique gene names and aliases. In contrast, Entrez Gene maintains 33024 human gene ids as of February 2006, consisting of 33006 unique official symbols. There are 51121 aliases, with 2438 shared by more than one gene. Overall, Entrez Gene maintains 80618 gene symbols and aliases. It is apparent that human gene references are not always obvious or consistent or stable.

Synonym ambiguity in literature

Usage of a unique symbol for each gene facilitates electronic information retrieval from publications and automated analysis (text mining). Unfortunately, historical usage of gene symbols in published literature is not fully captured by either HUGO or Entrez Gene. One example is the symbol AR. The University of Texas ARGH Biomedical Acronym Resolver [80] identifies 862 known acronym definitions in MEDLINE, while Stanford University Biomedical Abbreviation Server [81] identifies 1291 possible interpretations. AR is the official symbol for the androgen receptor gene (GeneID 367). A PubMed [82] search for AR produces 54,457 results. A search for 'androgen receptor' returns 9412, of which 6215 do not use the symbol AR.

Both HUGO and Entrez Gene record the usage of the symbol AR as an alias for one additional gene, AKR1B1 – more commonly called aldose reductase (GeneID 231). Entrez Gene also maps the alias AR to the AREG gene, or amphiregulin (GeneID 374). There are 425 PubMed results for amphiregulin, of which just 9 use AREG as a symbol, while 157 use AR. In addition, there are a number of other genes in PubMed using AR as an alias. One example is the 18-member adenergic receptor (adrenoceptor) gene family with 2671 PubMed results including AR as an alias. AR is often used in PubMed abstracts without a clear reference to a full gene name, requiring a domain expert for verification of the actual gene being described.

The absence of an automated approach for resolving ambiguity

between gene synonyms is a key problem [83, 84]. Natural Language Processing in particular is dependent upon term disambiguation, which has been called the "great open problem" of natural language lexical analysis [85]. In the biomedical domain, gene and protein name disambiguation is essential for providing quality protein-protein interactions, disease associations, and other complex biomedical analysis. This problem can also have a substantial impact on the efficiency of information retrieval methods, such as biomedical thesauri [86] or molecular pathway identification [87].

Resolving ambiguity

When dealing with gene name terminology, disambiguation tasks fall into two basic categories: identification of text which likely refers to a gene and, if so, which specific gene does the text address. As an example of the former, does "PI" refer to a gene (e.g. "glutathione S-transferase pi") or something else (e.g. "Permeability Index"). Once predicted as a gene name within a document, the symbol must be associated to a specific gene (e.g. distinguish between "glutathione S-transferase pi" and "serpin peptidase inhibitor"). Both of these issues can confound text analysis.

Automated disambiguation of gene and protein names can play a significant role in accelerating disease research and drug development. Natural language researchers began focusing on automated approaches to term disambiguation in the late 1980s and early 1990s. Yarowsky [88] used statistical models built from entries in Roget's thesaurus to assign sense to ambiguous words in text, using a Bayesian model to weight the importance of words related to the targeted ambiguous term. Gale, Church, and Yarowsky [89] outlined an approach that used the 50 words preceding and following the target term to define a context for that term's sense. In developing a method for general word sense disambiguation using unsupervised learning, Yarowsky [90] took a document classification approach to solving the problem of general term disambiguation. He showed in this study that generic English language terms often have only one sense per collocation with neighboring words.

Computational linguists and computational biologists have re-

cently begun to study term disambiguation in the biomedical domain. A number of researchers [85, 91] have proposed solutions that involve manually crafted rules to help natural language processing and information retrieval systems correctly process ambiguous synonyms. These rules are often combined with supervised learning methods (in which systems are provided with human-curated training data) and in some cases unsupervised learning methods (also often referred to as "clustering'). Recent work by Yu and Agichtein [92] compared four different approaches to solving the disambiguation problem: manual rules, fully supervised learning, partially supervised learning, and unsupervised. The manual method is then combined with several of the machine learning approaches to yield a system capable of extracting synonymous genes and proteins from biomedical literature. Liu et al. [93] also explored a partially supervised learning approach based on disambiguation rules defined in the Unified Medical Language System. In the case of both papers, results are promising, but the systems require a pre-existing set of handcrafted literature corpora (text sources), raising questions about scaling up to a level where a significant portion of human genes and proteins can be covered. Hatzivassiloglou et al. [94] applied machine learning to the problem of gene, protein and RNA molecule disambiguation in text, showing that accuracy levels, as defined by F-measure, of nearly 85% can be attained for classifying terms as belonging to the class of gene or protein. Research presented in paper IV of this thesis describes development of a large-scale human gene and protein name disambiguation system, seeking to correctly identify references to a specific gene in MEDLINE abstracts.

Faced with a question of how a compound modulates a pathway, a researcher is considering a complex problem. Issues to consider are adverse effects, pathological processes, toxicity, etc. Solution requires combined analysis of all available data, both literature and experimental, and a systems biology approach.

Present investigation

Imagination is more important than knowledge
Albert Einstein

This thesis presents a progression of studies related to the core hypothesis that bioinformatics approaches can facilitate the inference of functions for human protein-encoding genes.

Bioinformatics-based gene characterization, like all discovery processes, works best when multiple approaches lead to similar conclusions. The initial step is the selection of one or more genes for study. A predetermined or preexisting functional aspect of a gene or gene family, a gene's commercial application potential such as patentability, tissue specificity, subcellular location, or even the presence of a specific structural or functional domain can influence such selections. Second, the discovery process must be applied to a data collection or set of collections. Genome-scale experimental data, such as gene expression, can contain powerful clues as to the function of a protein. These clues are often lost without examination of specific results in light of additional, complementary information. As such, visualization tools are critical for discovery of relationships between genes in large-scale data. Beyond visualization, which relies on the accumulated knowledge of an individual, the final step is to draw on the accumulated knowledge in primary scientific literature. Detailed examination of scientific literature can thus link a gene with a role in a biological process.

Bioinformatics methods can facilitate and accelerate the discovery of gene characteristics. The investigation presented in this thesis involves development of methods and bioinformatics applications that assist in identification of poorly characterized genes and aid characterization efforts to elucidate their function.

- Prediction of the level of characterization of human genes or gene families, allowing selection of poorly understood genes for functional studies

- Identification of protein domains with unknown function, yet conserved across species throughout evolution
- Exploratory visualization of complementary genomic-scale, experimental and computational data with parallel heatmaps
- An automated text classification system for large-scale disambiguation of human gene names in literature

These methods and tools provide essential resources for targeted or systematic functional gene characterization efforts.

Paper I

Gene Characterization Index: A Metric for Assessing How Well We Understand Our Genes

In genome sequencing, progress can be measured simply by enumerating the number of sequenced base pairs. For gene annotation, however, there is no quantitative measure against which the progress of characterization efforts can be assessed.

The present paper introduces the first Gene Characterization Index (GCI), a computational method for scoring the extent to which each protein-encoding gene is functionally described. Inherently a reflection of human perception in a window of time, GCI serves both as a tool for identifying those genes which are least (or most) well characterized and for the assessment of annotation progress on the genome scale. The GCI scoring method was created based on the results from a global survey of life science researchers, who assigned characterization scores ranging from one (poor) to ten (complete) for a sample of genes.

Using the survey results as training data, machine learning methods were applied to develop a scoring function to assign scores to all human protein-encoding genes. The Entrez Gene and EnsEMBL databases were utilized to obtain quantitative gene annotation characteristics for use as potential model attributes. The final set of attributes from which the model could select consisted of: single nucleotide polymorphisms (SNPs); number of DNA sequences available in GenBank and EnsEMBL; InterPro protein domains; Gene Ontology annotations; KEGG metabolic pathways; OMIM disease associations; annotations in PRINTS, PROSITE, RefSeq and SwissProt; PDB protein structures; HomoloGene similar sequences in other species; descriptive annotations such as HUGO gene name and symbol and functional description from Entrez Gene; and PubMed literature references from Entrez Gene.

The performances of numerous classification algorithms were compared, including linear models (LM), regression trees (RT), neural networks (NN), support vector machines (SVM), and multivari-

ate additive regression splines (MARS). Most of the methods performed in a comparable range, suggesting that performance is not constrained by the statistical methodology. We selected a MARS model, as the MARS procedure provides the greatest clarity on which data attributes are used to optimize the fit to training data. The final model utilized the number of GenBank DNA sequences, InterPro domains, KEGG pathway associations, PubMed abstracts and OMIM disease references, and number of isoforms present in the SwissProt database.

Based on gene annotations archived in past releases of the GeneLynx database, the GCI scoring procedure was applied to study the temporal changes in gene characterization (Figure 5A), to identify poorly and well characterized genes within pharmaceutically relevant protein classes (Figure5B), and to highlight poorly characterized genes for which gene-specific patent applications exist from which potentially useful annotations could be obtained. Current GCI scores for all human genes are available on the GCI website. GCI scores have been provided for the human genes in the NovelFam3000 database (paper II).

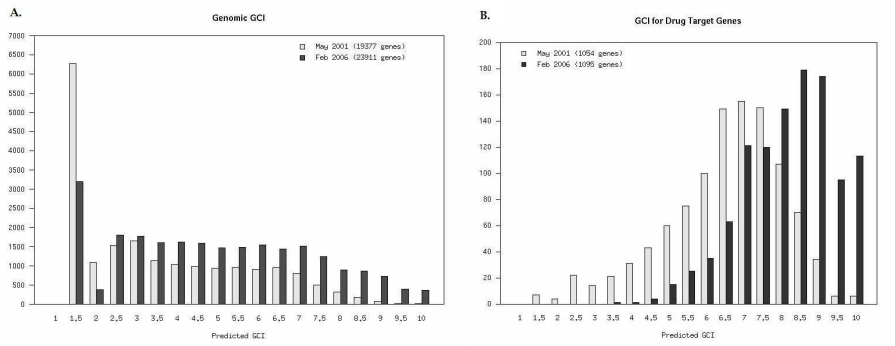

Figure 5. Histograms of GCI scores for human genes. Comparison of GCI scores for 19,377 genes with cDNA sequences from the GeneLynx database in May 2001 (light gray bars) and 23,911 genes from Entrez Gene in February 2006 (dark gray bars). A. Full genome scores show an general increase in the level of characterization and nearly 50% reduction in the number of completely uncharacterized genes. B. A subset of human genes representing drug target genes from the Drug Target database include 1054 genes from May 2001 (light gray bars) and 1095 genes from February 2006 (dark gray bars). The median GCI score has increased from 6 to 8.5, with close to 10% scored at a maximum value of 10.

The existing GCI scores reflect the state of life scientists' opinions at the time of the survey. As science advances, perspectives will change - our expectations for the depth of gene annotations will likely more stringent. In addition, over time, the available types and sources of data change. Therefore, the GCI scoring function must be periodically updated to reflect the available annotation resources and the changing opinions of researchers. The GCI website facilitates community input on gene characterization scores. As this feedback is collected and after the quality is evaluated, new generations of GCI scoring functions can be developed.

Author Roles: Raf Podowski was involved in survey preparation and validation, model development, testing, implementation and validation, and preparation of all results. Danielle Kemmer assisted in the survey development and training data preparation. Jochen Brumm provided guidance on statistics, classification and regression algorithms. Claes Wahlestedt assisted in the project design and application potential. Boris Lenhard assisted in obtaining gene attributes from historic GeneLynx databases. Wyeth Wasserman conceived the gene characterization index, assisted in the survey development, model attribute selection and performance validation, as well as use case identification. Raf Podowski, Danielle Kemmer and Wyeth Wasserman drafted the manuscript.

Paper II

NovelFam3000 – Uncharacterized Human Protein Domains Conserved Across Model Organisms

Approaches combining both experimental and bioinformatics methods and data may elucidate functional characteristics of uncharacterized protein domains.

The foundation of this project is based on identification of uncharacterized protein domains in human genes. Significance of a domain is indicated by its conservation across evolution, especially if found in model organisms such as yeast, worm or fly (Figure 6), thus providing broader informatics and experimental resources for functional studies. The final product of the project is a data centre, NovelFam3000, serving as a central resource for annotation of uncharacterized domain-containing proteins. The system provides access to dispersed Internet resources containing gene-specific experimental data, and allows for addition of annotation comments and posting relevant experimental results.

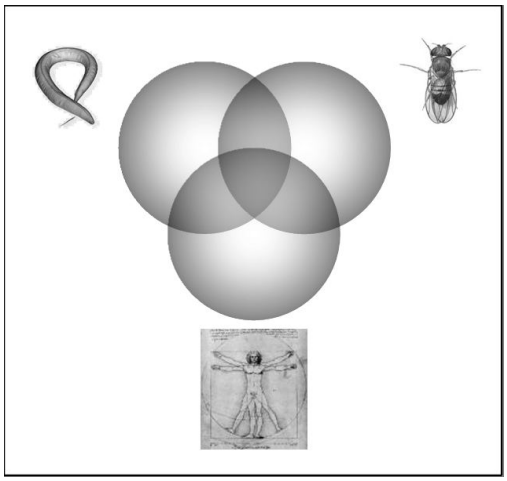

Figure 6. Project Objective. The goal of the project is the identification of uncharacterized human protein domains conserved across model organisms including D. melanogaster and C. elegans.

The data centre contains approximately 3000 protein domain families with minimal available biochemical annotation. Analysis of the Pfam database's domain families from Pfam-B and Domains of Unknown Function (DUFs) identified uncharacterized protein domains represented in each of three metazoan genomes, those of human, worm and fly (Figure 7). The selected protein domain families were required to have multiple human protein members.

An up-to-date GCI score (paper I) is displayed for each human

gene in the system. The characterization index allows users to observe changes in the functional understanding of an individual gene or whole domain family over time, or to identify uncharacterized domains in well-characterized genes for further study.

Consistent experimental results between multiple members of a domain family allow for inferences of the domain's functional role. We unite bioinformatics resources and experimental data in order to accelerate the functional characterization of scarcely annotated domain families.

Author Roles: Danielle Kemmer participated in the design of the study and generated experimental data. Raf Podowski participated in the compilation of a collection of novel-domain containing proteins, developed the Gene Characterization Index and contributed to the database design. David Arenillas and Jonathan Lim carried out the database development. Emily Hodges and Peggy Roth contributed experimental data for model organisms. Christer Höög coordinated the generation of the experimental data for the database. Erik Sonnhammer supervised the initial compilation of a collection of novel-domain containing proteins. Wyeth Wasserman conceived of the NovalFam3000 database and assisted in the interface design. Danielle Kemmer, Christer Höög, and Wyeth Wasserman drafted the manuscript.

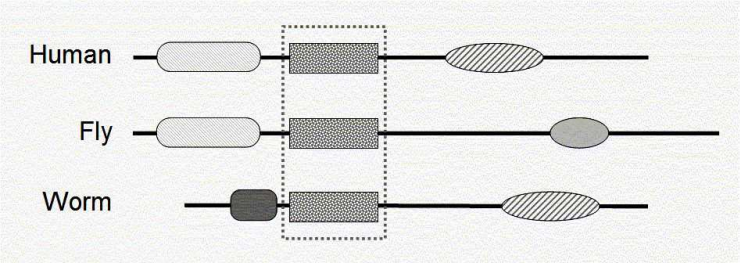

Figure 7. Conserved domain identification. Alignment of protein sequences for 3 species identifies a previously unidentified, conserved domain with unknown function. Clues to its function may come from new information about any single member of the group.

Paper III

Visualization of Complementary Systems Biology Data with Parallel Heatmaps.

Interpretation of large-scale biological data can be aided by the use of appropriate visualization tools. Heatmaps have emerged as a preferred technique for the display of genomics data, as they provide an extra dimension of information in a two dimensional display. However, an increasing focus on the integration of data from multiple sources for gene characterization (paper I and II) has created a need for the display of additional dimensions.

The application described in the present paper was developed to enable biologists to visually compare multiple gene-centric data sources, facilitating the discovery of significant functional relationships between genes and characteristics. Examples of such gene-centric data classes include: gene expression profiles, binding sites for transcription factors (TFs), Gene Ontology terms, disease/pathway annotations, literature-based associations (paper IV) and sub-cellular localization (paper II). A thorough search and examination of existing tools failed to identify a sufficiently flexible or ready solution. In order to endow researchers with the capacity to seek correlations between such disparate classes of data, we developed the Parallel HeatMap (PHM) viewer for four-dimensional data display. The flexible data entry structure of the Parallel HeatMap viewer facilitates the display of both continuous and discrete data. Confidence can be built by directly comparing computational predictions to experimental results (Figure 8).

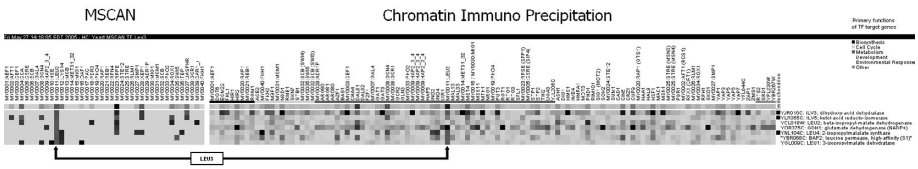

Figure 8. Correlations between in vivo and in silico binding sites. There is strong agreement between computational predictions generated by the MSCAN software using distinct yeast binding profiles of the Lue3 transcription factor and microarray-assessed chromatin immuno-precipitation results ("ChIP on Chip").

The Parallel HeatMap viewer enables knowledgeable life science researchers to observe patterns and properties within high-throughput genomics data in order to rapidly identify biologically logical relationships.

Author Roles: Raf Podowski initiated the project, participated in the development of functional requirements, design and testing of the parallel heatmap viewer, collected and generated data sources used in the validation, performed testing and result generation, drafted the manuscript and prepared the project web page. Brett Miller created the GenePilot software and multi-platform installation packages. Wyeth Wasserman provided guidance in identifying areas which may benefit from application of a parallel visualization approach, identified references and sources of data used in validation, ascertained results of all studies and revised the manuscript.

Paper IV

SureGene, a Scalable System for Automated Term Disambiguation of Gene and Protein Names

TEXT MINING IS FAST becoming a key enabling technology in drug discovery and an increasingly prominent topic at conferences. Uncoordinated selection of names for genes has created a significant problem for the automated analysis of biomedical literature. Automated disambiguation of gene and protein names could significantly help improve the efficiency of text analytics in the biomedical domain, accelerating disease research and drug development. Researchers are hindered by a lack of standard naming conventions for genes and proteins and must thus endure long, and sometimes fruitless, literature searches. Over 20,000 human genes have been identi-

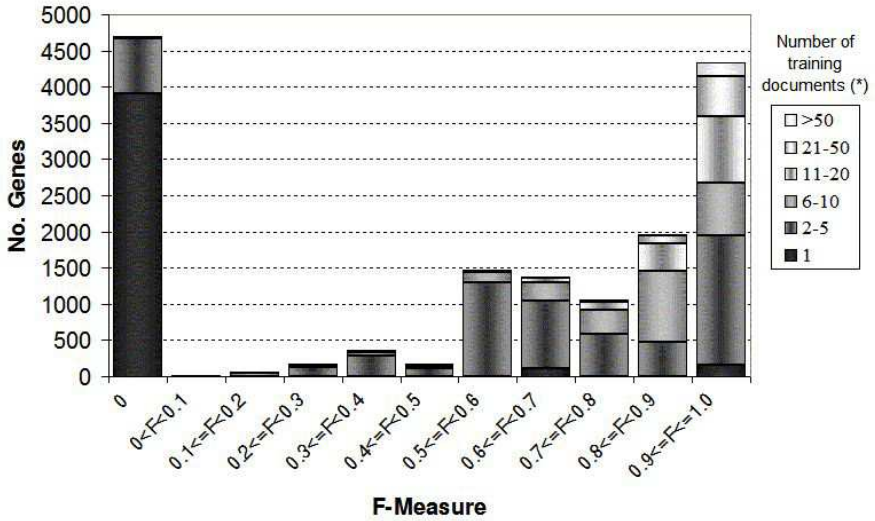

Figure 9. Effect of number of training document on gene context predictive accuracy. It can be seen that almost all the poorer performing gene models have small numbers of training documents. In addition, more than 84% of gene models with more than five training documents have an accuracy of greater than 70%. For models with more than 10 training documents, this increases to 91% of the models.

fied in Entrez Gene (formerly LocusLink) and over 100,000 different names have been used to refer to them. Automated disambiguation of gene and protein names can play a significant role in identification of articles with new or unique gene characterization information for improved gene annotation (Paper I and II).

In this paper, we present SureGene, a system for performing automated term disambiguation that can easily scale to tens of thousands of unique gene and protein names. SureGene uses a combination of machine learning and natural language processing technologies to identify abstracts relevant to specific genes and return these results as a ranked list. The SureGene system is able to automatically assign gene names to their Entrez Gene IDs in previously unseen MEDLINE abstracts.

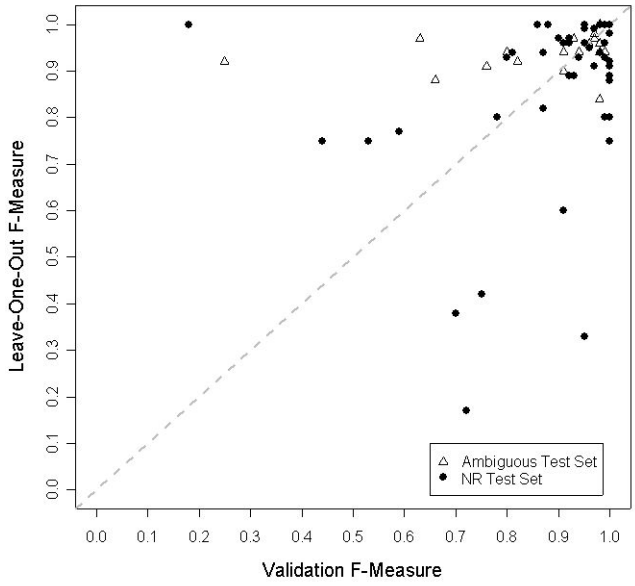

Figure 10. Validation of Gene Models Performance. F-Measure comparisons between model predictive performance, based on a Leave-One-Out (LOO) method, and real-life predictive performance for a set of 20 genes with highly ambiguous gene symbols and 46 genes from the human NR gene family. Points below the diagonal indicate models where human validation showed higher performance than that suggested by the LOO evaluation. Points above the diagonal reflect instances where the classification system's LOO estimate is more optimistic than the results obtained by human validation.

We show that SureGene is capable of accurately distinguishing between highly ambiguous gene terms, as well as between synonymous gene and non-gene terms (Figure 9). Two gene sets were selected for the purpose of real-world data validation. The first set consists of genes with known ambiguity in their gene names. A second set consists of Nuclear Receptor family genes. The results agreed well with the automated assessment. Accuracy levels for these genes, as defined by F-measure, were mostly 90% or higher (Figure 10).

We conclude that it is possible to achieve high quality gene disambiguation using scaleable automated techniques. Such disambiguation is important for searching and as a basis for other text analytics. Natural Language Processing in particular is dependent upon gene/protein name disambiguation to provide quality protein-protein interactions, disease associations, and other complex biomedical analysis.

Author Roles: Raf Podowski participated in idea development and project planning, data acquisition, model development, testing and validation, results generation and presentation, as well as manuscript preparation. John Cleary assisted in methodology development and provided guidance and practical assistance in application of the classification algorithms. Nicholas Goncharoff played a central role in project coordination, planning, system implementation requirements and manuscript preparation. Gregory Amoutzias performed validation of the system for nuclear receptor gene publications. William Hayes played a central role in the project conception, design of methodologies and vision for the practical implementation of the system.

Concluding remarks/perspectives

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

Isaac Asimov

A multi-front approach to gene characterization

Gene characterization moves most rapidly when multiple sources of information provide similar insights into the function or role of a protein. Therefore we need to draw on diverse data sources, including text and high-throughput genomics studies such as provided by microarrays to understand genes. The tools presented in this thesis provide a foundation that can be built upon to construct a more integrated analysis approach.

A vision for the future

A critical aspect of gene characterization for the future is selection of candidate genes within an area of interest that require expanded study. Eventually I envision a unified software system that takes as input a biological or medical subject and returns a list of genes likely to be involved. Such a system would integrate diverse forms of gene-centric data and would further highlight where deficiencies preclude or hinder successful analysis.

At the center of such an integrated system will be text analysis. The primary literature remains the fertile source of inspiration in cell biology and medicine. As the algorithms for text mining improve, the vast untapped literature will become accessible for inference.

While bioinformatics is central to this vision of the future, the fundamental discoveries remain tied to the industrious work of laboratory scientists. Bioinformatics methods should be grounded in this reality, seeking to provide researchers with assistance in exploring directions rather than stockpiling lists of predictions.

Acknowledgements

The road I took to arrive at the finish line of this doctoral thesis has been a journey through a labyrinth. With a couple of initial false starts, drastic course changes as well as great life events over a period of more than 10 years. All the research constituting the main body of the thesis has been performed in the last four years under the guidance of Wyeth Wasserman, while working full time at AstraZeneca and Oracle. Two of the projects, disambiguation and visualization, derived directly from needs of researchers at AstraZeneca. During all this time, I have had the pleasure of working with some excellent teams and individuals in both academia and industry. Although I wish to express my thanks to all of them, I give special thanks to the following individuals:

Wyeth Wasserman for going well out of his way to give me the opportunity to complete my doctoral research, despite the geographic distance and overflowing plate of other tasks and responsibilities. Always full of ideas and advice, with an excellent attitude. Best supervisor a student could wish for.

William Hayes for unfading support, encouragement and faith in my ability to produce quality results, for helping me become more critical and systematic in my work, for loyal friendship and tons of home improvement advice. I look forward to another chance to work together.

Charlie Berger for being a supportive and motivating boss during my time at Oracle, allowing me to develop some sorely needed presentation and management skills. Although I still prefer Python over PowerPoint.

Kasian Franks for excellent collaboration, continued support, loyal friendship, vision and unyielding direction, drive and perseverance.

Brett Miller for making the ideas of data visualization a reality and being a loyal and supportive friend.

Claes Wahlestedt, as co-supervisor for continued support of my academic efforts over a long and turbulent period.

Nicko Goncharoff for great support and encouragement in the SureGene project, excellent writing and editing skills and friendship.

Matti Nikkola for invaluable assistance in getting all academic records and administrative details sorted out making the conclusion of this work possible.

Lenore Clesceri from Rensselaer Polytechnic Institute for support and encouragement, as well as letting a senior engineering student try some molecular biology experiments. Her influence has been a major driving factor to this day.

At the **Center for Molecular Medicine and Therapeutics** at the **University of British Columbia** in Vancouver: **Jochen Brumm** for helping me get started with R and sharing a few pints with me in Vancouver; **David Arenillas** and **Dimas Yusuf** for great assistance on the Gene Characterization Project in the realm of databases and web design; **Dora Pak** for invaluable assistance with communications and logistics.

Facuty, staff and friends at the **Center for Genomics and Bioinformatics**, especially **Danielle Kemmer** for all the collaborative efforts, advice and assistance in projects. Also **Boris Lenhard** for helping me sort out the insides of the GeneLynx database, **Eugene Zabarovsky** and **Erik Sonnhammer** for research supervision and guidance, and **Björn Andersson** for support and academic guidance.

Faculty and fellow students at **Uppsala University**, especially **Charles Kurland** and **Siv Andersson** for allowing an engineer to try his hands in the lab, **Alireza Zomorodipour** for constant willingness to advise and instruct me in experimental methods, and **Thomas Sicheritz-Ponten** for getting me hooked on bioinformatics, the Python programming language and The Discworld. Oook!

Many friends have encouraged me over the years to keep going. I thank them all. Especially **Lisa Meijer** for being a delightful friend and setting a good example through perseverance in her own doctoral quest, **Erik Möller** for speaking his mind and many good times, **Eva Mauritzsson** for continuous curiosity, hospitality and encouragement, and **William Krivan** for being a great and positive individual exemplifying how to combine many diverse aspects of life.

Many a time, my focus needed realignment. The primary treatment derived from disciplined physical training. I wish to thank my

teachers and fellow practitioners of Uechi-ryu karate in Lexington, Massachusetts. Your example and dedication has helped me maintain strength and focus in all aspects of life. Special thanks to Sensei **Alan Azoff** and **Steve DiOrio**, **Jen**, **Matt**, **Steven** and **Scott**.

Finally, I would like to thank my family. Extra thanks to my parents and brother for continued encouragement and support. Very special thanks to my children, **Max** and **Bella**, who graciously accepted their father's need to work all evening way too often. Finally, extra special thanks to **Camilla**, my wife, for lasting support and putting up with an ever positive, yet dynamic and elusive finishing deadline.

References

1. Southan C. Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics* **4**, 1712–1726 (2004).
2. Orchard S, Hermjakob H, Apweiler R. Annotating the human proteome. *Mol Cell Proteomics* (2005).
3. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
4. Bogue, M.A. & Grubb, S.C. The Mouse Phenome Project. *Genetica* **122**, 71–74 (2004).
5. Mashimo, T., Voigt, B., Kuramoto, T. & Serikawa, T. Rat Phenome Project: the untapped potential of existing rat strains. *J Appl Physiol* **98**, 371–379 (2005).
6. Rual, J.F. et al. Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library. *Genome Res* **14**, 2162–2168 (2004).
7. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002).
8. Pennisi, E. Human genome. A low number wins the GeneSweep Pool. *Science* **300**, 1484 (2003).
9. Gewin V. A golden age of brain exploration. *PLoS Biol* **3**, e24 (2005).
10. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54–D58 (2005).
11. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (1997). (http://www. genecards.org)
12. SwissProt (http://www.expasy.org/sprot/)
13. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens P, De Smet F, Tranchevent L, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. Gene prioritization through genomic data fusion. *Nature Biotechnology* **24**, 537–544 (2006).
14. Stuart JM, Segal E, Koller D, Kim SK. A Gene Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* **21**, 21 (2003).
15. Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R. SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Research* **11**, 1574–1583 (2001).
16. Kemmer D, Huang Y, Shah SP, Lim J, Brumm J, Yuen MM, Ling J, Xu T, Wasserman WW, Ouellette BF: Ulysses – an application for the projection of molecular interactions across species. *Genome Biology* **6**, R106 (2005).
17. Shannon W, Culverhouse R, Duncan J. Analyzing microarray data using cluster analysis. *Pharmacogenomics* **4**, 41–52 (2003).
18. Robinson MD, Grigull J, Mohammad N, Hughes TR. FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* **3**, 35 (2002).

19. Huges, TR. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–26 (2000).
20. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* **28**, 37–40 (2000).
21. Moreau Y., De Smet F., Thijs G., Marchal K., De Moor B. Functional bioinformatics of microarray data : from expression to regulation. *Proceedings of the IEEE* **90**, 1722–1743 (2002).
22. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology* **4**, R28 (2003).
23. Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).
24. Hosack DA, Dennis G Jr., Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biology* **4**, R70 (2003).
25. Medical Subject Headings (http://www.nlm.nih.gov/mesh/)
26. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C and Eddy SR. The Pfam Protein Families Database. *Nucleic Acids Research,* Database Issue **32**, D138–D141 (2004).
27. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, and Wasserman WW. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Research* **33**, 3154–64 (2005).
28. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research* 2001, **11**, 2120–2126 (2001).
29. van Noort V, Snel B, Huynen MA. Predicting gene function by conserved co-expression. *Trends Genet.* **19**, 238–242 (2003).
30. Stuart JM, Segal E, Koller D, Kim SK: A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
31. Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF. Atlas – a data warehouse for integrative bioinformatics. *BMC Bioinformatics* **21**, 6:34 (2005).
32. HomoloGene (www.ncbi.nlm.nih.gov/HomoloGene/)
33. Kemmer D, Huang Y, Shah SP, Lim J, Brumm J, Yuen MM, Ling J, Xu T, Wasserman WW, Ouellette BF. Ulysses – an application for the projection of molecular interactions across species. *Genome Biology* **6**, R106 (2005).
34. Meyer RD and Cook D. Visualization of data. *Current Opinion in Biotechnology* **11**, 89–96 (2000).
35. IEEE Computer Society (www.computer.org)
36. ACM (www.acm.org)

37. International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV) (2001) (http://csdl.computer.org/comp/proceedings/cmv/2004/2179/00/2179toc.htm)
38. Boukhelifa N, Rodgers PJ. A model and software system for coordinated and multiple views in exploratory visualization. *Information Visualization* **2**, 258–269 (2003).
39. Ross G, Morrison A, Chalmers M. Coordinating Views for Data Visualization and Algorithmic Profiling. *Second International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'04)*, 3–14 (2004).
40. Bertini E, Catarci T, Kimani S, Santucci G. Exploiting Multiple Views to Support Visual Exploration and Mining. *Second International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'04)*, 15–23 (2004).
41. Munzner T, Guimbretiere F, Tasiran S, Zhang L, and Zhou Y. TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. *SIGGRAPH 2003, published as ACM Transactions on Graphics* **22**, 453–462 (2003).
42. Slack J, Hildebrand K, Munzner T, and St. John K. SequenceJuxtaposer: Fluid Navigation For Large-Scale Sequence Comparison In Context. *Proc. German Conference on Bioinformatics*, 37–42 (2004).
43. Breitkreutz BJ, Stark C, Tyers M. Osprey: a network visualization system. *Genome Biology* **4**, R22 (2003).
44. Altermann E, Klaenhammer TR. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics* **6**, 60 (2005).
45. KEGG: Kyoto Encyclopedia of Genes and Genomes. (http://www.genome.jp/kegg)
46. Reactome. Cold Spring Harbor Laboratory, European Bioinformatics Institute, and GO Consortium. (http://www.reactome.org)
47. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **11**, 2498–504 (2003).
48. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr., Koh KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL et al. An Information-Intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
49. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster Analysis and Display of Genome-wide Expression Patterns. *Proc. Natl. Acad. Sci.* **95**, 14863–14868 (1998).
50. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO. Genomic Expression Responses to DNA-damaging Agents and the Regulatory Role of the Yeast ATR Homolog Mec1p. *Molecular Biology of the Cell* **12**, 2987–3003 (2001).

51. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Baker JL, Somogyi R. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci.* **95**, 334–339 (1998).
52. Hierarchical Clustering Explorer (http://www.cs.umd.edu/hcil/hce/index.html)
53. Seo J, Shneiderman B. Interactively Exploring Hierarchical Clustering Results. *IEEE Computer* **35**, 80–86 (2002).
54. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).
55. Wu W, Noble WS. Genomic Data Visualization on the Web. *Bioinformatics* **20**, 1804–1805 (2004).
56. ArraySCOUT (LION Bioscience AG, Heidelberg, Germany)
57. GeneSpring (Silicon Genetics, Redwood City, CA, USA)
58. Spotfire DecisionSite for Functional Genomics (Spotfire, Goeteborg, Sweden)
59. Gana Dresen IM, Hüsing J, Kruse E, Boes T, Jöckel K-H. Software Packages for Quantitative Microarray-Based Gene Expression Analysis. *Current Pharmaceutical Biotechnology* **4**, 417–437 (2003).
60. Lee SG, Lee WS, Kim YS/ GOODIES: GO Based Data Mining Tool for Characteristic Attribute Interpretation on a Group of Biological Entities. *Genome Informatics* **14**, 675–676 (2003).
61. Nishimura K, Abe K, Ishikawa S, Tsutsumi S, Hirota K, Aburatani H, Hirose M. A PCA Based Method of Gene Expression Visual Analysis. *Genome Informatics* **14**, 346–347 (2003).
62. Segal E, Taskar B, Gasch A, Friedman N, Koller D. Rich probabilistic models for gene expression. *Bioinformatics* **17**, 243S–252S (2001).
63. Davidson GS, Hendrickson B, Johnson DK, Meyers CE, Wylie BN. Knowledge Mining With VxInsight: Discovery Through Interaction. *Journal of Intelligent Information Systems* **11**, 259–285 (1998).
64. Keim D, Kriegel HP. VisDB: Database Exploration Using Multidimensional Visualization. *IEEE Computer Graphics and Applications* **14**, 40–49 (1994).
65. Swanson DR and Smallheiser NR. Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neurosci Res Comm* **15**, 1–9 (1994).
66. Masys D. Linking microarray data to the literature. *Nature Genet* **28**, 9–10 (2001).
67. Glenisson P., Coessens B., Van Vooren S., Mathys J., Moreau Y., De Moor B. TXTGate : Profiling gene groups with text-based information. *Genome Biology* **5**, R43.1–R43.12 (2004).
68. MEDLINE (http://www.nlm.nih.gov/pubs/factsheets/medline.html)
69. Yahoo! Inc. (http://www.yahoo.com)
70. van Rijsbergen CJ. *Information retrieval*. 2nd ed. London: Buttersworth (1989).

71. Kowalski G. *Information retrieval systems: Theory and implementation.* Norwell, MA: Kluwer Academic (1997).
72. Buckley C and Lewit AF. Optimizations of inverted vector searches. In *Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 97–110. New York: ACM Press (1985).
73. Cutting DR, Karger DR, Pedersen JO, and Tukey JW. Scatter/Gather: A cluster-based approach to browsing large document collections. *SIGIR: 1992*, 318–29 (1992).
74. Zamir O, Etzioni O, Madani Om and Karp RM. Fast and intuitive clustering of Web documents. In *Knowledge, discovery, and data mining*, 287–90 (1997).
75. Kubat M, Bratko I, Michalski RS. A review of machine learning methods. In *Machine learning and data mining methods and applications,* ed. Kubat M, Bratko I, Michalski RS, **3**. New York: Wiley (1997).
76. Gennari JH, Langley P, Fisher D. Models of incremental concept formation. *Artif Intell* 40, 11–61 (1989).
77. Iliopoulos I, Enright A, Ouzounis C. Textquest: Document clustering of Medline abstracts for concept discovery in molecular biology. *Pacific Symposium on Biocomputing,* 384–95 (2001).
78. Jones G, Robertson AM, Santimetvirul C, Willett P. Non-hierarchic document clustering using a genetic algorithm. *Information Research* **1**, (1995).
79. Human Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/)
80. The University of Texas ARGH Biomedical Acronym Resolver (http://invention.swmed.edu/argh/)
81. Stanford University Biomedical Abbreviation Server (http://abbreviation.stanford.edu)
82. PubMed (http://www.pubmed.gov)
83. Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assoc* **9**, 621–636 (2003).
84. JS. What big pharma wants. *Genome Tech* **29**, 31–38 (2003).
85. Resnik P, Yarowsky D. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Nat Lang Engi* **5**, 113–133 (2000).
86. Aronson AR, *Ambiguity in the UMLS Metathesaurus*, National Library of Medicine (2001).
87. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: A machine learning approach. *Bioinformatics* **1**, 1–10 (2001).
88. Yarowsky D. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities* **34**, 179–86 (2000).
89. Gale WA, Church KW, Yarowsky D. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* **26**, 415–39 (1992).

90. Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* 189–96. Cambridge, MA (1995).
91. Rindflesch TC, Aronson AR. Ambiguity resolution with mapping freetext to the UMLS metathesaurus. *J Am Med Inform Assn* 240–4 (1994).
92. Yu H, Agichtein E. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics* ISMB Supplement, 340–49 (2003).
93. Liu HF, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assn* **9**, 621–36 (2002).
94. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: A machine learning approach. *Bioinformatics* **17**, S97–106 (2001).