Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 37 (2004) 43–53

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data

Andrey Rzhetsky,[a,b,c,*] Ivan Iossifov,[a,b] Tomohiro Koike,[e] Michael Krauthammer,[b] Pauline Kra,[b] Mitzi Morris,[a] Hong Yu,[d] Pablo Ariel Duboué,[d] Wubin Weng,[d] W. John Wilbur,[f] Vasileios Hatzivassiloglou,[d] and Carol Friedman[b]

[a] *Columbia Genome Center, Columbia University, New York, NY 10032, USA*
[b] *Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA*
[c] *Center for Computational Biology and Bioinformatics (C_2B^2), Columbia University, New York, NY 10032, USA*
[d] *Department of Computer Science, Columbia University, New York, NY 10032, USA*
[e] *Hitachi Software Engineering Co., Ltd., Yokohama, Japan*
[f] *National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20984, USA*

Received 13 August 2003

Abstract

The immense growth in the volume of research literature and experimental data in the field of molecular biology calls for efficient automatic methods to capture and store information. In recent years, several groups have worked on specific problems in this area, such as automated selection of articles pertinent to molecular biology, or automated extraction of information using natural-language processing, information visualization, and generation of specialized knowledge bases for molecular biology. GeneWays is an integrated system that combines several such subtasks. It analyzes interactions between molecular substances, drawing on multiple sources of information to infer a consensus view of molecular networks. GeneWays is designed as an open platform, allowing researchers to query, review, and critique stored information.

Keywords: Text mining; Bioinformatics; Information extraction; Molecular networks; Molecular interactions; Database; Artificial intelligence; Knowledge engineering; Machine learning

1. Introduction

Imagine a tribe of bright, but ignorant, cavepeople trying to understand the operation of a modern car by analyzing a collection of damaged cars produced by various makers. After many hours of hard manual labor, the cavepeople disassemble the cars into myriad small parts. Some are damaged, whereas some are intact. Some pairs of pieces interact with each other, whereas others do not interact. Some pieces are different in different cars, yet apparently have the same function. The leap to understanding the whole from knowing the parts requires reduction of redundant or conflicting pieces of information to a consistent consensus model that can be used for dynamics analysis. Researchers in the field of molecular biology of the post-genome era are in a situation similar to that of the junkyard cavepeople, save that they are contemplating a collection of diverse pieces of cellular machinery. Complicating the researchers' horizon, the identical piece of cellular machinery may play different roles in different cells of the same organism, or even within the same cell but under different environmental conditions, just as a Swiss Army knife in the car glove compartment can be used for cutting wood, sewing fabric, or removing a cork from a bottle under appropriate circumstances. The number of nodes in human molecular networks is measured in hundreds of thousands when all substances (genes, RNAs, proteins, and other molecules) are considered together. These numerous substances can be in turn present or absent in dozens of cell types in

* Corresponding author. Fax: 1-212-851-5149.
E-mail address: ar345@columbia.edu (A. Rzhetsky).

1532-0464/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2003.10.001

humans—clearly, the complexity is too great to yield to manual analysis. Thus, with the hope of relieving the information overload currently assaulting scientists, we are developing GeneWays, a computer system that integrates a battery of tools for automatic gathering and processing of knowledge on molecular pathways.

Various components of the GeneWays system were described in the earlier publications; the present paper provides a synthesis and overview of the project as a whole, indicating interactions between system modules and directions of the planned future development of the system.

2. Background

It would be impossible to give a complete review of the vast area spanning text analysis and molecular interactions databases, even were we to allow this review to consume the page limit of this article. Nevertheless, it is important to give at least a cursory overview of recent accomplishments and key research areas related to the work described in the current paper. These key research topics correspond to the major computational problems encountered by a researcher on her long and winding road from a collection of plain-English texts to a useful database of molecular interactions.

2.1. Document sorting

First, given a large database of abstracts of journal articles, such as PubMed (http://www3.ncbi.nlm.nih.gov/Entrez/index.html), the researcher needs to distinguish papers relevant to her interests from millions of non-relevant ones. For example, she might be interested in articles having "cell cycle" in the title or abstract and less interested in articles talking about supercolliders or fur export. This task is *document sorting*; it can be viewed as a classical task of machine learning, the problem of how to do automated classification of objects into two or more classes—"relevant" and "non-relevant," in this case. Such classification can start with a set of examples provided with known class assignment (*supervised* machine-learning methods) or without such a training dataset (*unsupervised* learning) [1–3]. The implemented unsupervised approaches to document sorting include clustering of article abstracts [4], assuming that relevant and non-relevant clusters are likely to form separate groups. Supervised methods applied to this problem include naïve Bayes classifier [5,6], and support-vector machines [7,8].

2.2. Term identification

Second, given a set of documents that she believes is relevant to her interests, the researcher needs to *identify terms* [9], such as names of genes, proteins, diseases, and tissues. Term identification is a critical text preprocessing stage required by many natural-language processing engines, including GENIES [10]. Researchers attempted to attack this problem by inferring morphological rules that guide generation of a term [11,12], by using parts-of-speech tagging engines that can help the downstream applications to identify multiword noun phrases [11,13], grammar rules [14], combinations of rule-based and dictionary-based methods [15], support-vector machines [16], hidden Markov models [17], and naïve Bayes and decision-trees classifiers [18]. It appears that the problem of tagging biological terms is a difficult one, and that we may achieve better results by combining several of these approaches. Early approaches [11,13] were tested on small test sets and reported excellent reports where the precision and recall were over 90%. However more recent results reported for larger test sets achieved results that ranged in the mid 70s and 80s [14,19,20] for precision and recall.

2.3. Term meaning disambiguation

Third, having identified terms, our researcher realizes that the problem of term identification is confounded when a term has multiple meanings (*term ambiguity*), and when multiple terms correspond to the same concept (*term synonymy*). For example, the name *p21* can refer to a gene, a protein, or a messenger RNA, depending on the sentence context. Deducing the right meaning is known as *sense disambiguation*, a problem that can be tackled with machine-learning approaches, such as those using naïve Bayes, decision trees, or inductive-learning classifiers [21]. The most common examples of the synonymous names are pairs of abbreviated and complete protein names (e.g., *il2* stands for *interleukin-2*; both terms often occur in biological texts). The problem of synonyms can be alleviated with automatically generated dictionaries [22–24]. See Liu 2002 [24] for an overview of word sense disambiguation applied to the biomedical domain.

2.4. Information extraction

Fourth, once she has identified and disambiguated such terms, our researcher wants to do *information extraction* (remember that our researcher wants to extract information about molecular interactions). She has her choice of methods that vary in complexity and success. The first group of approaches are "correlation methods" that exploit information about co-occurrence of terms in articles or abstracts [1,25–27]. In a more sophisticated form, such methods are based on a hidden Markov model [28] that requires no dictionary of terms. Methods of the second group target information extraction via *template matching*: they identify regular

expressions in the text using a term dictionary and a collection of hand-crafted patterns [29–32]. Methods of the third group explicitly use *formal grammars* that can identify nested structures in a sentence. In a nutshell, a grammar is a set of allowed symbols (usually divided into terminal words that we observe in a sentence and non-terminal, invisible symbols that serve as intermediates in the imaginary process of generating a sentence), and a set of production rules that have the capability of expressing not only regular expressions but also nested structures. Production rules are used to generate a sentence by stepwise substitution, starting with a single top-level non-terminal symbol, and ending with a sentence that contains only terminal words. Given a grammar and a valid sentence, we can reconstruct the sequence of substitution events (usually expressed as a *parse tree*) leading to generation of the sentence by the grammar; this process is called *parsing*. The GeneWays project as well two other molecular-biology-related linguistic projects [10,33,34] use grammar-based parsers. Since different projects have different foci and are typically tested on small datasets, it is currently impossible to tell with confidence what is the relative performance of these methods, although we expect that grammar-based methods have a higher precision. A grammar-based method, however, requires access to a dictionary listing properties of the words it recognizes (the *lexicon*) and information about allowable combinations or patterns of words that are encoded in its rules. Such information is currently supplied by manual analysis of sample texts in consultation with domain experts.

2.5. Ontology

Fifth, imagine that our researcher has struggled through a multiplicity of research articles and has managed to extract a large number of statements; she now needs to store this information in a database. Therefore, she requires a knowledge model on which to build a database schema. Various knowledge models for molecular-biology data have been suggested over the past few years, and many of them have been implemented in databases; all these databases (except the GeneWays database) were created manually (see [35] for a review). The most famous projects of this type include the EcoCyc/MetaCyc knowledge base and ontology, the primary emphasis of which is bacterial pathways [36,37], the Gene Ontology of sequence/structure conservation across eukaryotes [38], the Tambis Ontology [39], the Ontology of Molecular Biology [40], the ontology for conceptual modeling of biological information [41], and RiboWeb: the ontology/database of structural models of the ribosome [42]. There are also various databases of molecular interactions that have an implicit ontology: KEGG, LIGAND [43,44] (databases of diverse molecular interactions and protein–ligand interactions, respectively); BIND [45], DIP [46], and MINT [47] (three databases of protein–protein interactions); BindingDB [48] (a knowledge base of diverse molecular interactions and associated affinity information), and COMPEL [49] (a compendium of protein–DNA interactions). The GeneWays project is also provided with a knowledge model [50] that is fine-tuned for analysis of signal-transduction pathways in eukaryotes, but can be used for representing bacterial data as well.

2.6. Visualization

Sixth, thinking of summary of molecular interactions as a blueprint of a computer chip (a real computer chip is usually less complex than a living cell), our researcher certainly needs to visualize fragments of the map to get insights into mechanisms of the ''chip's'' work; graph drawing is a large field in its own right. An excellent review of available methods related to molecular biology is provided by [51]; a general treatment of graph drawing problems can be found in book by Di Battista et al. [52].

2.7. Integrated system

Seventh, rather than straggling with individual tools every time she needs to process a new batch of a few thousand articles, the researcher may decide to integrate the previous six computational steps in a single system. The GeneWays system described in this proposal is just such an integrated system. Similar systems include the PIES system in Singapore, developed for analysis of protein–protein interactions described in journal abstracts [31,53–55], the GENIA system in Japan [56] that uses knowledge extraction from both article abstracts and full articles to cross-index those articles with Internet-based databases, and the United States—developed MEDSTRACT system [32,57] that extracts relationships of the form ''A inhibits B'' from journal-article abstracts.

We have set the context for our own project, GeneWays, by covering briefly the work that other groups have done on molecular pathways and on automated analysis of research articles. We have been developing the GeneWays system for 5 years at Columbia University; we recently used it for analysis of nearly 150,000 full-text articles, and as a result were able to populate a prototype database with nearly 1.5 million unique statements. We believe that GeneWays is a state-of-art system that can be considerably extended and enhanced, and that can be used as a tool for exciting research projects.

3. GeneWays: motivation and anatomy

The word ''GeneWays'' probably emerged from an aberrant fusion of words ''genes'' and ''pathways.'' The

system was designed with the ambitious goal of automating extraction of information on molecular interactions locked in the text of journal articles.

Since the potential scope of the term ''information on molecular interactions'' is immense, at the first phase of the system development, we decided to focus on molecular interactions pertinent to signal-transduction pathways. Although the division of molecular pathways into ''metabolic'' and ''signal transduction'' is probably just a convenient way of looking at the elements of an interconnected unified system, there are distinct differences between these two types of pathways. Metabolic pathways mostly deal with tremendously diverse chemical alterations of relatively small molecules, whereas signal-transduction pathways are relatively poor in chemical mechanisms and predominantly involve ''switch-on'' and ''switch-off'' interactions among large molecules, such as genes and proteins. In an article describing a signal-transduction pathway statements that ''protein A binds protein B,'' ''protein C phosphorylates protein D,'' and ''protein E activates gene F'' are seen frequently, although gene and protein names (A–F in the current example) can be drawn from a sizable list of nearly a hundred thousand names. Signal-transduction pathways, therefore, seem to be an easier target for information extraction from free text, although soon after starting the project we realized that even this ''easier'' task is extremely difficult to perform correctly.

GeneWays is designed to extract relations (or actions as we call them in our ontology; see [50]) between substances or processes. If we think about pathways as oriented graphs, we can divide relations into two groups: direct and indirect. Direct relations, which usually are physical interactions between substances, correspond to a single edge in the graph; indirect relations link two nodes (substances or processes) with a series of two or more edges. Direct relations in the current version of GeneWays include *N-acylate*, *N-glycosylate*, *O-glycosylate*, *acetylate*, *attach* (= *bind*), *createbond*, *degrade*, *demethylate*, *dephosphorylate*, *breakbond*, *methylate*, *overexpress*, *phosphorylate*, *express*, *contain*, *transcribe*, *release*, *interact*, and *substitute*. Indirect relations (which occasionally can also correspond to direct relations) include *activate*, *actupon*, *cause*, *generate*, *inactivate*, *limit*, *promote*, and *signal*. The GeneWays database currently maintains the following concept types: complex, disease, domain, gene, geneoprotein, process, protein, species, and smallmolecule.

Only a subset of these concepts (gene, geneoprotein, process, protein, and smallmolecule) can serve as vertices of a pathway graph; GeneWays uses the remainder to capture additional information about defined vertices and edges of the oriented graph. (We are currently implementing the additional concepts described in [50].)

Here, we describe two views of GeneWays: from the perspectives of a system developer and of a user.

From the point of view of a developer, GeneWays looks as in Fig. 1 (stars identify modules of the system that are developed but not yet integrated). We can think of a system as an engine that processes raw data to create a structured product. The ''raw data'' that come

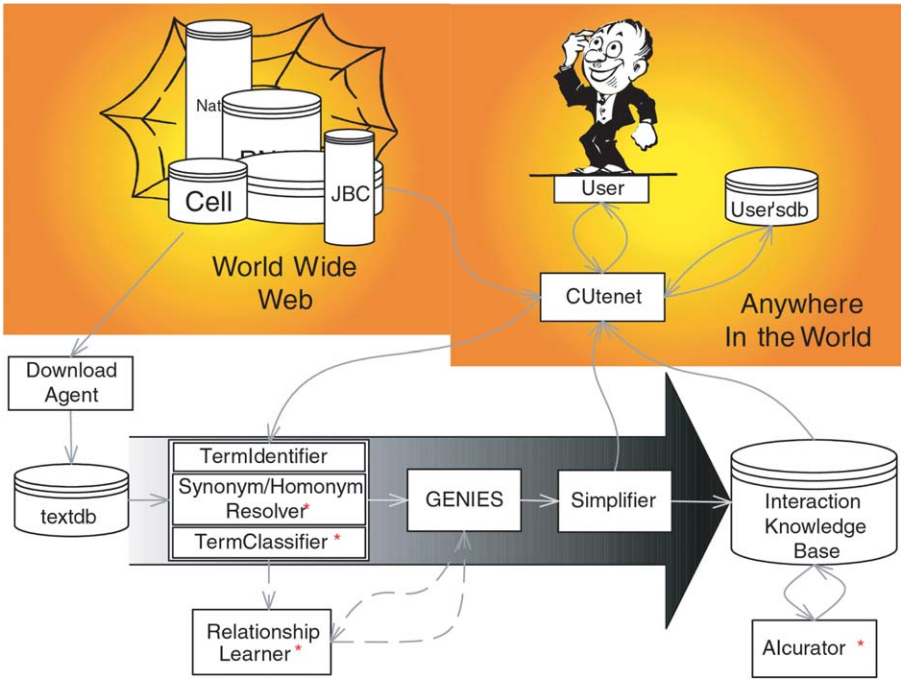

Fig. 1. A simplified view of GeneWays system.

into the system are represented by electronic copies of research articles coming from the World Wide Web, such as from the web sites of scientific journals for which developers have a legitimate subscription. The task of collection and local accommodation of numerous research articles (we have approximately 150,000 full-text articles in the current system) is done by a GeneWays module called the "Download Agent," which saves retrieved text into a local database, as shown in Fig. 1. The heart of the system comprises the modules shown inside the big arrow in Fig. 1. First, the Term Identifier module [19] identifies biologically important concepts in the text, such as names of genes, proteins, processes, small molecules, and diseases. Many such terms have synonyms and homonyms, so the Synonym/Homonym Resolver module clarifies the meaning by assigning a "canonical" name to each concept multiple aliases. Furthermore, there are other kinds of ambiguity associated with terms. For example, term "interleukin-2" can identify the corresponding gene, a messenger RNA, or the protein, depending on context. The Term Classifier module ([21]) resolves sense ambiguity of this type. GENIES is a natural-language processing parser [10] that takes as input plain text with identified and tagged concepts (for example term "interleukin-2" can be tagged as "<substance = 'p'>interleukin-2< \substance>," where 'p' stands for "protein"). The output of GENIES is represented with semantic trees that are not intended to be directly comprehended by humans, because they represent complex nested relationships captured from text in a machine-readable form.

An example of a GENIES parsing is shown in Fig. 2.

The Simplifier module takes these complex output trees and unwinds them into simple binary statements (an example of a simple binary statement is "interleukin-2 binds interleukin-2 receptor"—the statement links two substances, interleukin-2 and interleukin-2 receptor, with action "bind"). The resulting simplified statements are saved into the Interaction Knowledge Base, which is the main resource associated with the GeneWays system. The Interaction Knowledge Base is implemented on the basis of a commercial relational database (Oracle 9i), and is built on GeneWays ontology [50].

Note that the automatically generated knowledge base is of necessity noisy: the GeneWays system extracts some percentage of statements incorrectly, and, even among correctly extracted statements, we should expect redundancy and contradictions. Therefore, the database requires curation, a process in which the original statements are annotated with statements regarding confidence in the corresponding information. The traditional way to perform such curation is through manual labor of human experts—a monumental task even for the database at its current size of roughly 3 million redundant statements extracted from 150,000 articles. To reduce the manual work, we are implementing a Curator module that would allow GeneWays to compute the estimates of reliability automatically. We recently suggested a plausible approach to the curation and annotation problem, and we are in the process of implementing it [58].

```
[action, promote,
        [geneorprotein, mdm2],
        [action, degrade,
                [process, ubiquitin proteolytic pathway],
                [geneorprotein, p53]
        ],
]
```

Fig. 2. The GENIES parsing of the sentence. Recent studies have reported that mdm2 promotes the rapid degradation of p53 through the ubiquitin proteolytic pathway.

The two remaining modules are the CUtenet and Relationship Learner. We describe the first, CUtenet, later when explaining the user's perspective. The Relationship Learner module has a unique role within GeneWays because its relationships with other modules (shown by dashed lines in Fig. 1) is different from the other relationships in the system. Most of the relationships in the figure (shown by solid arrows) depict flow of information during the data processing that leads to populating the Knowledge Base. The Relation Learner module works with the output of Term Identification/Disambiguation module to identify new semantic patterns that developers can use later to improve GENIES; therefore, the arrows connecting the Relationship Learner module with the rest of the system depict information flow during system-improvement cycles, rather than during data-processing cycles.

From the point of view of a user, the system is represented by its portal, CUtenet (pronounced "See-u-tenet," which stands for "Columbia University tenet," or "cute net," whichever you prefer; see [59], a standalone program that accesses both the Knowledge Base and the GeneWays pipeline, as directed by a user. The primary function of CUtenet is visualization of user-defined pathways. Recently we augmented the program to access the GeneWays Interactions Knowledge Base, to retrieve various interactions defined by a query formulated by a user, and to visualize these interactions on the monitor. Moreover, the user can request information about the sentences corresponding to individual interactions and even can see the full articles from which the sentences were extracted. (Each interaction in the database is linked to a full-text article stored in the publisher's web site. The users of GeneWays system would be able to see the full-text article only if they have a legitimate subscription to the corresponding journal.) As an illustration of how the system works, let us consider the following example. Imagine that you are interested in a substance, the protein called collagen. You are formulating a query equivalent to a question "Show me all interactions for collagen." The total number of

interactions for a single substance stored in the Knowledge Base can be overwhelming (for collagen it is more than a thousand) and you need some mechanisms for reducing complexity of the output. Since each relation is frequently captured by GeneWays more than once from different sentences in the same as well as in distinct articles, the simplest filter for reducing the complexity of CUtenet figures is the number of times that each relation is entered into the Knowledge Base from independent sentences. In the case of collagen, the requirement that the interaction with collagen occur in the database at least 15 times retrieves collection of only 12 interactions (Fig. 3A). Reduction of threshold to 10, 5, and 0 repetitions brings about 25, 74, and 1335 interactions, respectively (see Figs. 3B–D, respectively). Clearly, it is not very useful to show all 1335 interactions available for collagen at once. We certainly realize that this simple filter is imperfect because the statements repeated more frequently are not necessarily more important or more reliable than those repeated less frequently, nevertheless this simple filter is better than no filter at all. We are developing a set of sophisticated filters that will allow users to select intuitive concepts for choosing among statements, such as the probability of a statement being true (see [58]). In the current version of GeneWays a user can "walk" through the database by requesting that she visualize interactions for substances that are already shown on the screen (Fig. 3E). Furthermore, by clicking on a graph edge in CUtenet window the user can retrieve original sentences

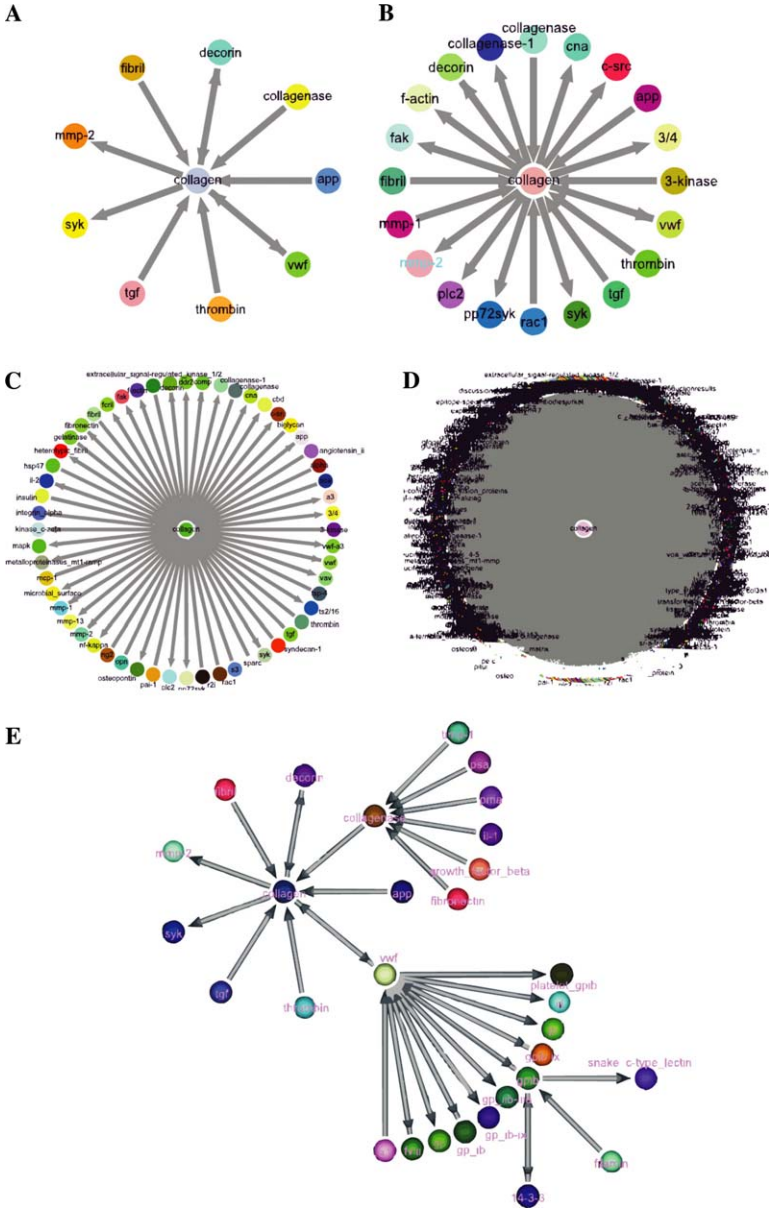

Fig. 3. Examples of output of queries of Interaction Knowledge Base visualized with CUtenet module.

corresponding to the interaction and full articles containing these sentences (see Fig. 4).

An alternative way to access GeneWays system through CUtenet is shown in Fig. 1; here, the user submits a request for processing of her favorite journal article through the GeneWays pipeline; such processing is culminated with a visualization of the extracted relationships. For example, by processing a plain-text version of a *Cell* article [60], GeneWays produced Fig. 5 which shows 46 interactions. This is a relatively high number: the average number of interactions extracted by GeneWays pipeline from an average *Cell* article is half this number. To obtain a more objective view of the number of interactions per article, we computed distributions of the number of statements per article extracted by GeneWays from three journals, *Cell*, *Journal of Molecular Biology*, and *Science*, from collection of articles spanning the past 5 years (Fig. 6). The average number of statements extracted from a single article in these three journals was 18.06, 20.72, and 4.33, respectively. These numbers may appear at first to be somewhat low especially for *Science* magazine; recall, however, that we analyzed all articles from each journal, and that *Science* publishes articles in all fields of science, rather than in only biology. It is natural, therefore, that an article on theoretical physics typically contains little information about interactions between genes and proteins.

4. Evaluation of extraction precision

One of important properties of a system is precision, defined as the ratio of the statements extracted correctly to the total number of extracted statements.

To evaluate the precision of GeneWays, we selected 2500 of the most frequent unique statements (out of several hundred thousand unique statements that are currently stored in GeneWays Knowledge Base). We then had an expert in molecular biology go through the 2500 list, checking correctness of extraction—the endeavor took a few weeks. According to this expert evaluation, 125 statements of the 2500 were either extracted with errors or corresponded to "phantom statements" generated by the GeneWays system. We then traced all stages of processing for each of these 125 statements, and found out that 100 of them were incorrect due to errors in term identification, 12 due to GENIES errors, and 5 due to Simplifier errors; 8 were actually correct (expert's error, as judged by the developers' team). Therefore, according to this evaluation, GeneWays' precision was 95%; GENIES recall was previously evaluated to be about 65% [10].

5. Current status of the system

The GeneWays system is far from being completely developed. For example, all modules marked with red asterisks in Fig. 1 are implemented in their prototype version, but have yet to be integrated with the GeneWays pipeline. It appears that the current precision bottleneck is associated with the term-identification module, which attempts to solve a formidable problem of recognizing biologically important terms in scientific publications; the problem appears to be harder when the terms are from biology rather than from medicine, business, or general English language [20]. We expect that our work on automated curation will give us

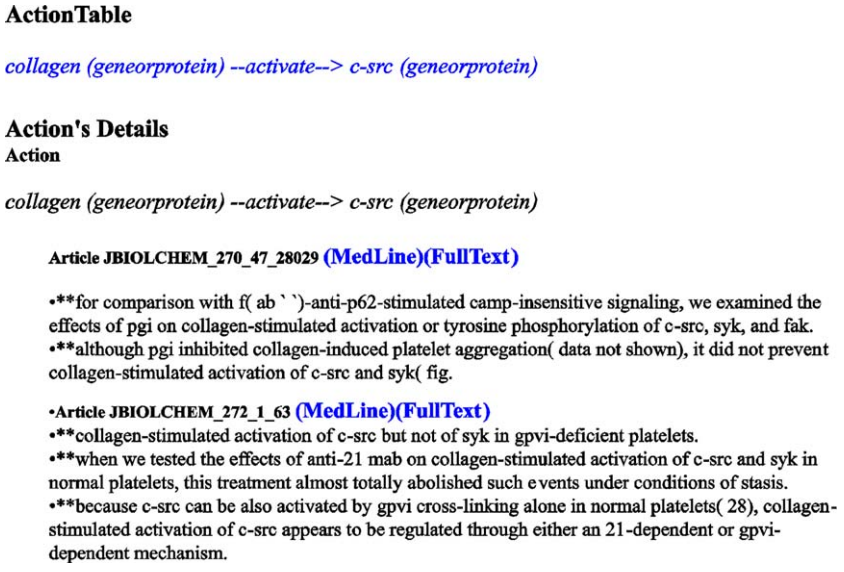

Fig. 4. A simplified view of information regarding interaction "collagen activates c-src" provided by GeneWays Knowledge Base. For each interaction visualized by CUtenet a user can obtain a list of sentences containing corresponding piece of information and complete articles containing these sentences.

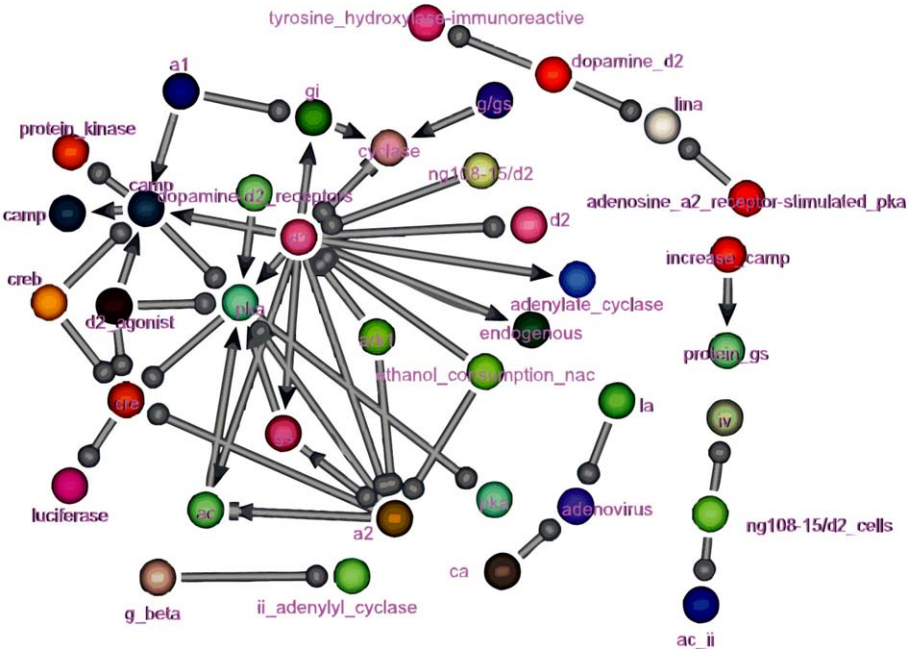

Fig. 5. Results of GeneWays analysis of a single *Cell* article. In this representation all binary relationships between molecules or processes are shown as oriented edges that end with either arrows, for "activate" relationships, or a ball, for all other relationships.

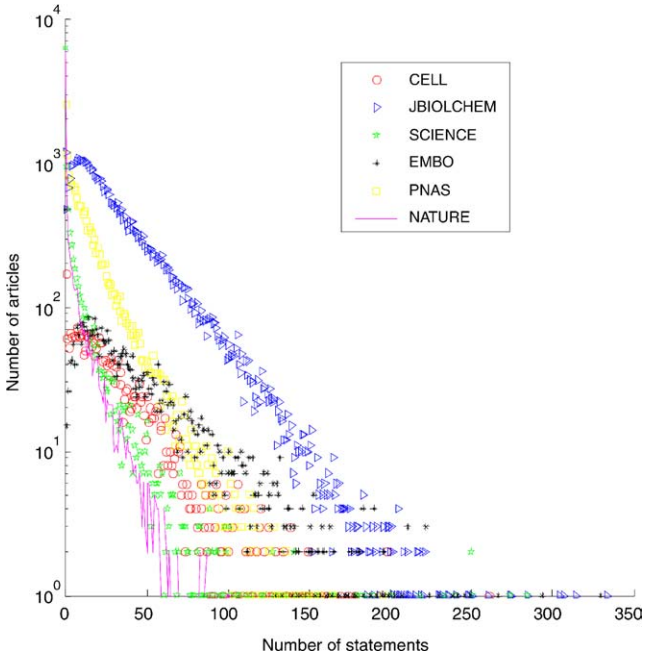

Fig. 6. Distribution of the number of statements extracted by GeneWays from a single journal article for six journals: *Science* (SCIENCE in the figure), *Cell* (CELL), *Journal of Biological Chemistry* (JBIOLCHEM), *Journal of European Molecular Biology Organization* (EMBO), *Proceedings of the National Academy of Sciences USA* (PNAS), and *Nature* (NATURE).

insights into the knowledge-generation work of scientific community at large, and also will point to ways we can improve our system. This work is likely to increase the value of the resulting database.

6. Discussion

6.1. Hand-made databases and automatically produced databases

There are a few popular molecular interaction databases, such as EcoCyc [61] and KEGG [44], that are populated by groups of careful experts. Such "manual" databases are designed to provide a consensus view of the evolving field of molecular biology (devoid of redundancy and inconsistencies), usually have a low error rate, and can express extremely complex statements about the underlying biological systems. In contrast, the GeneWays Knowledge Base is designed to capture a "stochastic" view of the field, where statements tend to repeat and conflict, and where each statement is associated with a publication time point. The GeneWays Knowledge Base is likely to include a larger number of errors than do the manual databases (note that, in general, rigorous evaluations of the precision of the manual databases are not undertaken), and the number of types of relationships extracted automatically is smaller than can be extracted by a human expert. However, automatic systems can populate quickly an extremely large database (much larger than our current database of 1.5 million unique statements), and repetitive conflicting statements extracted automatically can be treated essentially as experimental data (see [58]). Since the volume of text data currently available is tremendous, statistical approaches to analysis of statements extracted from the literature appear both promising and requisite.

6.2. Binary versus N-ary relationships

There is a contradiction between representation of pathways information optimized for convenience of computation and representation that captures text-encoded statement for optimized precision. Computation on a large pathway database is the easiest when all relationships are converted to a binary form, such as "protein A binds protein B." The text-encoded information often corresponds to not only binary, but also tertiary, quaternary, and higher-order relationships. For example, statement "proteins A and B synergistically activate gene C" represents a tertiary relationship that can be captured by GENIES (see [62] for detailed discussion of biomedical sublanguages). That relationship is simplified (broken down), for convenience of computation, into three binary statements: "protein A binds protein B," "protein A activates gene C," and "protein B activates gene C." Although computation is more efficient when statements are binary, the combined binary statements are not equivalent to the original tertiary relationship. The compromise that we have chosen is to keep in the knowledge base both representations, the binary and *N*-ary.

6.3. Nobody (and no system) is perfect

GeneWays in its current form has limitations. As follows from our evaluation of system precision, the noisiest part of the system is associated with term tagging (see [20] for a detailed discussion of this problem). In general, it is difficult to identify a name of substance or a process in a text: our favorite examples of difficult gene names include "forever young" (in plant *Arabidopsis thaliana*) and "mothers against decapentaplegic" (in fly *Drosophila melanogaster*). Improved term tagging is therefore likely to lead to a significant reduction in the error rate. Another plague crippling the system is associated with term synonymy. Although we compiled a database of gene/protein name synonyms, the dictionary approach alone appears to be insufficient. For example, "p53" and "p53 tumor suppressor" are currently stored in the GeneWays knowledge base as separate substances; these expressions have the same meaning but are difficult to recognize automatically as synonyms.

A totally different difficulty is associated with "translations" between sublanguages in scientific community. The same chemical event may be expressed in several strikingly different ways (different sublanguages) in different subdisciplines' research literature. For example, in the language of molecular biology, the statement "protein kinase A phosphorylates protein B" means the same thing as the expression

$$\mathrm{B} + \mathrm{ATP} \overset{\mathrm{A}}{\rightleftarrows} \mathrm{B} \cdot \mathrm{P} + \mathrm{ADP}$$

in the language of biochemistry (where ATP and ADP stand for adenosine triphosphate and adenosine diphosphate, respectively, and *P denotes a phosphate residue). Note that, in the biochemical description, kinase A is not part of the equation, but rather is merely a catalyst facilitating the reaction. A hard-core biochemist may argue that what molecular biologists say is incorrect; however, since both communities are able to understand their own statements correctly, we are dealing with two sublanguages requiring translation from one to another. If the articles analyzed by GeneWays are written in the language of molecular biology, but potential users of the resulting database speak in biochemical sublanguage (which is probably more precise), then automated "translation" of statements may become necessary.

6.4. Werewolves of biological terminology

There is a difficulty of recognizing terms "p53" and "p53 tumor suppressor" as synonyms—here the major problem is in deciding where protein name ends and a description of its function starts.

There are more extreme cases when a single term can be correctly interpreted in multiple ways. Our favorite example is protein name "MAPKKK," which stands for "mitogen-activated protein kinase kinase kinase."

Consider a hypothetical sentence "Mitogen-activated protein kinase kinase kinase phosphorylates protein Y." Term recognition here is a real problem because "mitogen," "mitogen-activated protein kinase," and "mitogen-activated protein kinase kinase" are valid substance names which are important for capturing pathway information contained in the sentence—sentence contains four interactions, namely "mitogen activates mitogen-activated protein kinase kinase kinase," "mitogen-activated protein kinase kinase kinase activates and phosphorylates mitogen-activated protein kinase kinase," "mitogen-activated protein kinase kinase activates and phosphorylates mitogen-activated protein kinase," and "mitogen-activated protein kinase kinase kinase phosphorylates protein Y."

6.5. Two types of redundancy in the database

The automatically generated GeneWays database has at least two sources of redundancy. One source is associated with redundancy of research literature: every statement viewed as important by a scientific community is repeated multiple times in various publications. By nature an image of the published information, the GeneWays knowledge base contains multiple instances of a large portion of the interactions represented in its database.

The second source of redundancy is less direct; it is associated with reasoning that can be done on the basis of a set of known molecular interactions. We mentioned

that all molecular interactions can be divided into two groups: "direct" and "indirect." For example, if protein A activates protein B by phosphorylation, and protein B activates gene C by binding to the promoter of gene C, interactions between A and B, and between B and C, are direct, whereas the interaction between A and C can be computed from direct interactions and is indirect. Since all indirect interactions can be deduced from direct ones *given that the set of direct interaction is complete*, we can conceive, and work to create a completely non-redundant database that contains only unique direct interactions. As our research field develops, certain direct interactions may become indirect, as the intermediate steps are discovered.

We conclude this paper by expressing what is perhaps the most powerful of the lessons that our work on GeneWays has taught us: that the field of analysis of biological and medical texts is replete with exciting unsolved problems, problems more than sufficient to entertain myriad of researchers for many decades.

Acknowledgments

The authors are grateful to Ms. Lyn Dupré and to two anonymous reviewers for valuable comments on the earlier version of this paper. This work was supported by Grants EIA-0121687, DE-FG02-01ER25500, and GM61372 from the National Science Foundation, Department of Energy, and National Institutes of Health, respectively.

References

[1] Shatkay H, Edwards S, Wilbur WJ, Boguski M. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. Proc Int Conf Intell Syst Mol Biol 2000;8:317–28.
[2] Shatkay H, Wilbur J. Finding themes in Medline documents: probabilistic similarity search. Ismb 2000.
[3] Shatkay H, Wilbur WJ. Finding themes in medline abstracts. In: IEEE Advances in Digital Libraries; 2000; 2000.
[4] Iliopoulos I, Enright AJ, Ouzounis CA. Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. Pac Symp Biocomput 2001:384–95.
[5] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol 1999:77–86.
[6] Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein–protein interactions. Bioinformatics 2001;17(4):359–63.
[7] Joachims T. A statistical learning model of text classification with support vector machines. In: Croft WB, Harper DJ, Kraft DH, Zobel J, editors. SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval. New York: ACM Press; 2001. p. 128–36.
[8] Joachims T. Transductive inference for text classifcation using support vector machines. In: International Conference on Machine Learning (ICML); 1999; 1999.
[9] Jacquemin C. Spotting and discovering terms through natural language processing. Cambridge, MA: MIT Press; 2001.
[10] Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;17(Suppl 1):S74–82.
[11] Fukuda K, Tamura A, Tsunoda T, Takagi T. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput 1998:707–18.
[12] Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2002;18(8):1124–32.
[13] Proux D, Rechenmann F, Julliard L, Pillet VV, Jacq B. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. Genome Inform Ser Workshop Genome Inform 1998;9:72–80.
[14] Gaizauskas R, Demetriou G, Humphreys K. Term recognition and classification in biological science journal articles. In: 2nd International Conference on Natural Language Processing (NLP-2000); 2000 June 4; Patras, Greece; 2000. p. 37–44.
[15] Rindflesch TC, Hunter L, Aronson AR. Mining molecular binding terminology from biomedical text. Proc AMIA Symp 1999;1:127–31.
[16] Kazama J, Makino T, Ohta Y, Tsujii J. Tuning support vector machines for biomedical named entity recognition. In: ACL-02. Philadelphia; 2002.
[17] Collier N, Nobata C, Tsujii J. Extracting the names of genes and gene products with a Hidden Markov model. In: Coling 2000. Germany: Saarbruken; 2000. p. 201–7.
[18] Nobota C, Collier N, Tsujii J. Automatic term identification and classification in biological texts. Proc Nat Lang Pac Rim Symp 1999;1999:369–74.
[19] Krauthammer M, Rzhetsky A, Morozov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. Gene 2000;259:245–52.
[20] Hirschman L, Morgan AA, Yeh AS. Rutabaga by any other name: extracting biological names. J Biomed Inform 2002;35(4):247–59.
[21] Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics 2001;17(Suppl 1):S97–106.
[22] Pustejovsky J, Castano J, Cochran B, Kotecki M, Morrell M. Automatic extraction of acronym-meaning pairs from MEDLINE databases. Medinfo 2001;10(Pt):371–5.
[23] Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc 2002;9(3):262–72.
[24] Liu H, Lussier YA, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. J Biomed Inform 2001;34(4):249–61.
[25] Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co- occurrences of gene names in Medline abstracts. Pac Symp Biocomput 2000:529–40.
[26] Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J. Detecting gene relations from Medline abstracts. Pac Symp Biocomput 2001:483–95.
[27] Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001;28(1):21–8.
[28] Leek RL. Information extraction using hidden markov models [Masters Thesis]. San Diego: University of California; 1997.
[29] Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein–protein interactions. Ismb 1999;1:60–7.
[30] Ono T, Hishigaki H, Tanigami A, Takagi T. Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics 2001;17(2):155–61.
[31] Ng SK, Wong M. Toward routine automatic pathway discovery from on-line scientific text abstracts. Genome Inform 1999: 104–12.

[32] Pustejovsky J, Castano J, Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput 2002:362–73.
[33] Park JC, Kim HS, Kim JJ. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. Pac Symp Biocomput 2001:396–407.
[34] Yakushiji A, Tateisi Y, Miyao Y, Tsujii J. Event extraction from biomedical papers using a full parser. Pac Symp Biocomput 2001;1:408–19.
[35] Stevens R, Goble CA, Bechofer S. Ontology-based knowledge representation for bioinformatics. Brief Bioinform 2000;1(4): 398–414 (9).
[36] Karp PD, Riley M, Saier M, et al. The EcoCyc Database. Nucleic Acids Res 2002;30(1):56–8.
[37] Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The MetaCyc Database. Nucleic Acids Res 2002;30(1):59–61.
[38] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25(1):25–9.
[39] Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. TAMBIS—transparent access to multiple bioinformatics information sources. Ismb 1998;6:25–34.
[40] Schulze-Kremer S. Ontologies for molecular biology. Pac Symp Biocomput 1998:695–706.
[41] Paton NW, Khan SA, Hayes A, et al. Conceptual modelling of genomic information. Bioinformatics 2000;16(6):548–57.
[42] Altman RB, Bada M, Chai XJ, Whirl Carillo M, Chen RO, Abernethy NF. RiboWeb: an ontology-based systems for collaborative molecular biology. IEEE Intell Syst 1999;14(5): 68–76.
[43] Goto S, Nishioka T, Kanehisa M. LIGAND: chemical database of enzyme reactions. Nucleic Acids Res 2000;28(1): 380–2.
[44] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27–30.
[45] Salama JJ, Donaldson I, Hogue CW. Automatic annotation of BIND molecular interactions from three-dimensional structures. Biopolymers 2001;61(2):111–20.
[46] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30(1):303–5.
[47] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett 2002;513(1):135–40.
[48] Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. Comb Chem High Throughput Screen 2001;4(8):719–25.
[49] Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, Wingender E, Kel AE. COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucleic Acids Res 2000;28(1):311–5.
[50] Rzhetsky A, Koike T, Kalachikov S, et al. A knowledge model for analysis and simulation of regulatory networks. Bioinformatics 2000;16:1120–8.
[51] Uetz P, Ideker T, Schwikowski B. Visualization and integration of protein-protein interactions. In: Golemis E, editor. Protein–protein interactions—a molecular cloning manual. Cold Spring: Cold Spring Harbor Laboratory Press; 2002. p. 623–46.
[52] Di Battista G, Eades P, Tamassia R, Tollis IG. Graph drawing. Algorithms for the visualization of graphs. Upper Saddle River, NJ: Prentice Hall; 1999.
[53] Wong L. PIES, a protein interaction extraction system. Pac Symp Biocomput 2001;1:520–31.
[54] Wong L. Kleisli, a functional query system. J Funct Programming 2000;10(1):19–56.
[55] Wong L. Bioinformatics integration simplified: the Kleisli way. In: Lai PS, Yap E, editors. Frontiers in human genetics: diseases and technologies. Singapore: World Scientific; 2001. p. 79–90.
[56] Collier N, Park HS, Ogata N, et al. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. EACL'99 1999;1:271–271.
[57] Pustejovsky J, Castano J, Sauri R, Rumshisky A, Zhang J, Luo W. Medstract: Creating Large-scale Information Servers for biomedical libraries. In: ACL-02; 2002; Philadelphia; 2002.
[58] Krauthammer M, Kra P, Iossifov I, et al. Of truth and pathways: chasing bits of information through myriads of articles. Bioinformatics 2002;18(Suppl 1):S249–57.
[59] Koike T, Rzhetsky A. A graphic editor for analyzing signal-transduction pathways. Gene 2000;259:235–44.
[60] Yao L, Arolfo MP, Dohrman DP, et al. betagamma dimers mediate synergy of dopamine D2 and adenosine A2 receptor-stimulated PKA signaling and regulate ethanol consumption. Cell 2002;109(6): 733–43.
[61] Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. The EcoCyc and MetaCyc databases. Nucleic Acids Res 2000;28(1):56–9.
[62] Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35(4):222–35.