Accomplishments and Challenges in Literature Data Mining for Biology

Lynette Hirschman
The MITRE Corporation

Jong C. Park
KAIST

Junichi Tsujii
University of Tokyo

Limsoon Wong
Labs for Information Technology

Cathy Wu
Georgetown University

Abstract

In this review, we summarize recent accomplishments in literature data mining for biology. We then discuss the need for a challenge evaluation for this field, and initial steps to create such an evaluation. Literature data mining has progressed from simple recognition of terms to extraction of interaction relationships from complex sentences, and has broadened from recognition of protein interactions to a range of problems such as improving homology search, identifying cellular location, or recognizing trends and themes in the literature. Given this explosion of research, we argue that the time is now right to create a challenge evaluation focused on one or more problems of immediate biological relevance. A challenge evaluation will give rise to a shared infrastructure, including annotated training and test data, and shared evaluation methods. This would enable researchers to compare approaches and share information, leading to accelerated progress in the field. In this context, we describe two specific applications: extraction of biological pathways from the literature and automated database curation. For each of these, we outline the task definition, the creation of an annotated corpus, and evaluation metrics.

1 Introduction

Even though the number and the size of sequence databases are growing rapidly, most new information relevant to biology research is still recorded as free text in journal articles and in comment fields of databases like the GenBank feature table annotations. As biomedical research enters the post-genome era, new kinds of databases that contain information beyond simple sequences are needed, for example, information on cellular localization, protein-protein interactions, gene regulation and the context of these interactions. The forerunners of such databases include KEGG [1], DIP [2], BIND [3], among others. They are still small in size and are largely hand curated. The development of reliable literature data mining technologies can accelerate their growth.

This paper reviews some recent accomplishments and applications of literature data mining technologies to biology. Section 2 reviews the work in terms of depth, and Section 3 shows the breadth of recent work. Most of the early work on automated understanding of biomedical papers concentrated on analytical tasks such as identifying protein names [4] or relied on simple techniques such as word co-occurrence [5] and pattern matching [6]. More recently, we have begun to see work based on more general natural language parsers that could handle considerably more complex sentences [7, 8]. This year, we see the emergence of sophisticated natural language technologies that can handle anaphora, as well as extracting a broader range of information [9, 10, 11, 12, 13, 14]. For the past three years, the Pacific Symposium on Biocomputing series has had a track dedicated to natural language processing and information extraction in biology. The response to the call for papers and the quality of the submitted papers mark this as an emerging field which combines bioinformatics and natural language processing in innovative and productive ways.

It is also apparent from these papers that there is no common yardstick for impartially assessing and fairly comparing the performance of these systems. We think it is important to the development of the field to organize a set of biologically significant challenge problems and to set up the corresponding evaluation benchmarks. We motivate this proposal by drawing from the experience in the newswire domain, where literature data mining techniques have been demonstrated to work well. Results from various evaluations show that information extraction systems can identify and classify names of person, organization, location, etc. at an accuracy of greater than 90%; and they can successfully extract binary relations among these entities, such as "ORGANIZATION located_at LOCATION" or "PERSON works_at ORGANIZATION" [16], at over 80% accuracy. In addition, other information access and retrieval techniques have proved effective in selecting documents relevant to a specific topic or to providing answers to questions based on information located in document collections: the leading systems can provide correct answers to factual queries at 75-85% accuracy [17].

Much of the progress in the newswire domain has come about because of systematic *common evaluations* conducted in natural language processing and information retrieval at the series of Message Understanding Conferences (MUC) [18, 19, 16] and Text REtrieval Conferences (TREC) [20, 17, 21]. Having a "challenge evaluation" for literature data mining in Biology can benefit biology in the way that the CASP evaluation[1] has accelerated progress in developing computational models for protein folding. The second half of this paper is devoted to this issue. Section 4 describes the goals of a challenge evaluation and the ingredients needed for a successful evaluation. Section 5 and Section 6 illustrate how a challenge evaluation can be set up in the areas of extraction of biological pathways and automated database curation.

2 Recent Accomplishments: Increasing Depth

We provide here a brief survey on extracting the interactions between proteins, drugs, and other molecules. The surveyed works illustrate the progress of the field and show the increasing depth of the natural language processing techniques that have been used.

Fukuda *et al.* [4] was a pioneer of the field. In this early paper, the focus was on identifying protein names. They encountered many protein names that were very long compound names; also, different names were used to identify the same protein, even within the same article; and furthermore, some protein names were also common English words. The solution proposed was to make use of special characteristics such as the occurrence of uppercase letters, numerals, and special endings to pinpoint protein names. While they did not report the precision of their method, a similar approach taken by a team at Molecular Connections Pvt Ltd of India suggested that a precision or specificity of over 70% was easily obtainable at an estimated recall or sensitivity of over 70%. The development of a large biology-specific corpus by Ohta *et al.* [22], together with techniques based on Hidden Markov Models [23] or on Bayesian classifiers trained on k-grams [24], would further raise both sensitivity and specificity in recognizing protein names.

Shortly after that, the field progressed rapidly beyond the problem of recognizing names and entered the realm of recognizing interactions between proteins and other molecules. The early works on this topic could be roughly divided into two main approaches. The first approach, represented by Stapley *et al.* [5] extracted co-occurrences of gene names from MEDLINE documents and used them to predict their connections based on their joint and individual occurrence statistics. This approach was subsequently followed up by Ding *et al.* [9], who systematically examined the impact on recall and precision of mining interaction information when an abstract, a sentence, or a phrase is used as the unit in which to check for term co-occurrence. The second approach, represented by Ng *et al.* [6], used templates that matched specific linguistic structures to recognize and extract protein interaction information from MEDLINE documents.

The pioneering work of Ng *et al.* was productively pursued by the use of natural language processing techniques of increasing sophistication. Wong [25] expanded the number of templates to increase sensitivity. Park *et al.* [7] introduced a bidirectional incremental parsing technique based on combinatory categorial grammar. Yakushiji *et al.* [8] used a full parser with a large-scale general-purpose grammar to analyze the argument structure in MEDLINE abstracts. While no extensive and careful validation results were available on these systems, their specificity was estimated in the 60%-80% range. All these papers could handle sentences whose structures were more complex than those handled by Ng *et al.* However, none of them could handle pronouns. And the overall sensitivity (recall) of these more complex systems remains an issue.

Most recently, works were reported that address the pronouns and co-referring sentences. Pustejovsky *et al.* [12] presented a robust parser for identifying and extracting inhibition relations from biomedical literature. The system uses a corpus-based approach to develop rules specific to a particular predicate or a

[1] Critical Assessment of Techniques for Protein Structure Prediction; see http://predictioncenter.llnl.gov/.

class of predicates. An interesting feature of this system was its anaphora resolution module. The results reported in this paper focused on *inhibition* relations and demonstrated that it was possible to extract limited but biologically important information from free text with high reliability using a classical natural language processing approach. Hahn *et al.* [10] described the MEDSYNDIKATE natural language processor designed for acquiring knowledge from medical reports. The system was capable of analyzing co-referring sentences and was also capable of extracting new concepts given a set of grammatical constructs. Leroy *et al.* [11] presented the GeneScene system whose parser used prepositions as entry points into phrases in the text, in contrast to the main trend which used verbs as entry points. It then filled in a set of basic templates of patterns of prepositions around verbs and nominalized verbs. It also had a set of rules for combining these templates to extract information from more complex sentences. Again, extensive and careful validation results were not available on these systems. However, based on small-scale experiments, these systems should have higher performance. For example, Pustejovsky *et al* reported that their system achieved 90% specificity at 57% sensitivity for extraction of inhibition relations.

3 Recent Accomplishments: Increasing Breadth

Besides the recognition of protein interactions from scientific text, natural language processing has also been applied to an increasingly broad range of information extraction problems in biology, such as identifying cellular localization, improving homology search, recognizing themes and trends, etc. This section provides a brief description of some recent applications of information extraction.

Baclawski *et al.* [26] described a diagrammatic knowledge representation technique called keynets. The rich ontology of the Unified Medical Language System was used to construct and index keynets. Using both domain-independent and domain-specific knowledge, keynets parsed texts and resolved references to construct new relationships between entities.

Humphreys *et al.* [27] described two information extraction applications in bioinformatics based on templates. The first application, EMPathIE, extracted details of enzyme and metabolic pathways from journal articles. The second application, PASTA, extracted information on the roles of amino acids and active sites in protein molecules from journal articles. This work illustrated the importance of template matching, and applied the technique to terminology recognition.

Rindflesch *et al.* [28] described EDGAR, a natural language processing system that extracted relationships between cancer-related drugs and genes from biomedical literature. EDGAR drew on a combination of technologies: a stochastic part of speech tagger, a syntactic parser able to produce partial parses, a rule-based system, and semantics information from the Unified Medical Language System. The metathesaurus and the lexicon in the knowledge base were used to identify the structure of noun phrases in MEDLINE texts.

Thomas *et al.* [29] presents the customization of an existing information extraction system called Highlight for the task of gathering data on protein interactions from MEDLINE abstracts. They developed and applied templates to every part of the texts and calculated the confidence for each match. The resulting system could provide a cost-effective means for populating a database of protein-protein interactions.

Chang *et al.* [30] modified the PSI-BLAST algorithm to use literature similarity in each iteration of its database search. They showed that supplementing sequence similarity with information from biomedical literature search could increase the accuracy of homology search result.

Iliopoulos *et al.* [31] described an algorithm for large-scale clustering of MEDLINE abstracts based on a statistical treatment of terms, together with stemming, a "go-list", and unsupervised machine learning. In spite of the minimal semantic analysis, clusters constructed in this paper provided a shallow description of the documents and supported concept discovery.

Stapley *et al.* [13] used a support vector machine to classify terms derived by standard term weighting techniques in order to predict the cellular location of proteins from their description in abstracts. The performance of the classifier on a benchmark of proteins with known cellular locations was better than that of a support vector machine trained on amino acid composition and was comparable to an expertly hand-crafted rule-based classifier [15].

Wilbur [14] formalized the idea of a "theme" in a collection of documents as a subset of the documents and a subset of the indexing terms such that each element of the latter had a high probability of occurring in all elements of the former. An algorithm was given to produce themes and to cluster documents according to these themes in an optimal way. The result of applying this method to over fifty thousand documents on AIDS was given as an illustration.

4 Organizing a Challenge Evaluation

The papers described in the preceding sections illustrate the enormous potential for the application of natural language processing and literature data mining techniques to biology. These techniques can be applied to the extraction of biological pathways, to the curation of biomedical databases, or to improving access to on-line literature. However, to date, few of these techniques have made it into routine use to help manage biological information.

To understand which techniques will work on which problems, we need to do systematic evaluation of these techniques. We propose that we can do this by defining biologically motivated common challenge problems that will attract a diverse set of researchers to solve these problems. A challenge problem would focus on a specific problem of biological importance; the organizers of the challenge problem would provide training data, blind test data and evaluation metrics. These would be presented to the research community, who would provide running systems that would be evaluated using a standard set of evaluation metrics. The participants would then meet, to discuss their experiences and learn from each other's experiences, to understand what worked and what didn't work.

We can identify the set of ingredients for a successful evaluation:

- Challenge problem. This should be a problem of biological significance, preferably one already being addressed, such as literature search to assemble biological pathways, or creation of specialized databases to organize information and make it accessible.
- Task definition. This defines the criteria for evaluation—what constitutes a "correct" answer in the context of the challenge problem. This requires a formal specification of the target output. For the biological pathway task, this might be a formal relational language consisting of a set of predicates (activate, inhibit,...), and classes of entities that can participate in that relationship (see Section 5 for examples). For the database curation task, it might consist of entries conforming to some specific ontology or nomenclature as specified by the database curators (see Section 6 for examples).
- Training data. To enable developers to build systems that solve the "challenge problem," developers need annotated data that can be used as "practice tests"—with the right answers provided. For a biological pathway, this might be a set of documents and the relations extracted from each of the articles. For a database curation task, it would be the database entries and the set of documents referenced in the curation. In addition, the training data must specify the *linkage* between the extracted information and the occurrences (phrases or sentences) in the associated article that provide the evidence for the extracted information. These linkages are the *annotations* that make it possible to create rules that map from the free text occurrence of information to the required target output.
- Test data. Once the system is built, it must be evaluated against *blind* test data—data that neither the system nor the developers have previously seen. This makes it possible to assess the generality of the solution. Note that the test data may not need to have the linkages annotated: it is sufficient to supply the input (free text) and target output.
- Evaluation methodology and implementation. The definition of a formal evaluation methodology is a key part of creating a challenge problem. There must be a reproducible method of evaluating system performance on the defined problem. Ideally, there would be an *automated* evaluation method and supporting software. This would allow participants to grade themselves on the training data (the "practice tests"); automated evaluation also supports system development techniques such as iterative hill climbing and machine learning.
- Evaluator. There must be a neutral group who runs the evaluation. The evaluator is responsible for providing the test data, for collecting the system runs on the test data, and for evaluating those runs.
- Participants. Any evaluation is only as good as the groups (and systems) that participate in it. Therefore, it is critical to identify beforehand a core set of groups who would be willing to perform such an evaluation, if the rest of the infrastructure were provided.
- Funding. To create a successful challenge evaluation, there must be funding for the infrastructure. The evaluation itself must be funded (in particular, the designated evaluator group), and finally, participants are more likely to participate if there is funding associated with the evaluation. For the participants, the association with funding may be indirect, e.g., it may be sufficient that there are funded programs (government or private) that might directly or indirectly reward good results in such an evaluation.

We next look at two examples of challenge problems: the extraction of biological pathways from the literature and techniques for automating database curation.

5 Extraction of Biological Pathways

To a biologist, a biological pathway is generally a chain of events and decision points that pertain to specific biological functions, such as the production of a desaturated fatty acid. In contrast to the situation with genes, where a detailed ontology for their classification and annotation has been established [32], there is no widely accepted ontology for biological pathways. Moreover, the ideal ontology that most closely reflects biological pathways as a biologist sees them may not be suitable for information extraction tasks. Here, we adopt a more relaxed view instead and consider biological pathways as a network of interactions and events between proteins, drugs, and other molecules. We propose three layers of challenges with respect to this more relaxed view. At the first layer, the task is to recognize names of proteins, drugs, and other molecules. At the second layer, the task is to recognize basic interaction events between molecules. At the third layer, the task is to recognize the relationships between the basic interaction events.

Before we describe the three tasks above in more detail, let us first set up the framework for benchmarking these tasks. The framework is oriented towards information extraction rather than deep natural language understanding. That is, we see each task as filling in a set of prescribed templates for each problem, as opposed to obtaining detailed parse trees and complete semantic representations of each sentence. We have three reasons for this orientation. First, filling in a template is closer to the application scenario of filling in a database table. Second, information extraction may not necessarily be syntax-based, and hence the present choice allows us to assess a broader range of techniques. Third, we do not have grammars that provide sufficiently broad coverage for the language found in the biomedical literature. This is because the language is complex, parsers are not yet that robust, and because the articles or abstracts may not always be written in grammatical English.

The framework is as follows. A number of test databases are constructed. Each test database is organized as a set of records. Each record should have a piece of text to be tested and a list of expected facts. The text can be at the sentence level, the abstract level, or the entire article level. The list of expected facts should contain everything that a "perfect" information extractor for the task on hand can extract and nothing else. For convenience, each fact can be thought of as a short sentence in a highly standardized form such as "P_1 activate P_2". More abstractly, we see a test database db as a set $\{(t_1, F_1), ..., (t_m, F_m)\}$, where t_i are the texts and $F_i = \{f_{i,1}, ..., f_{i,n_i}\}$ are expected facts. There are two primary levels of evaluation. The first is at the level of individual records. The second is at the level of the entire test database.

The traditional performance evaluation of information retrieval systems calls for the following. At either level, we evaluate the sensitivity (or recall) and specificity (or precision) of an information extractor E against the list of expected facts, where

$$recall(E) = \frac{TP(E)}{TP(E) + FN(E)} \qquad precision(E) = \frac{TP(E)}{TP(E) + FP(E)}$$

The definitions for $TP(E)$ (ie. true positives), $FN(E)$ (ie. false negatives), and $FP(E)$ (ie. false positives) depend on whether we are evaluating at the record level or at the database level. Note that it is not possible to define the usual notion of true negatives in our context because there is no theoretical bound on the number of "facts" that can be generated from a sentence and because it is not reasonable to use the closed world assumption in biology. At the record level, each expected fact in a separate record is counted as a separate instance. If $E(t)$ is the set of facts that E extracts from a text t, then

$$TP(E) = \sum_{(t,F)\in db} |E(t) \cap F| \quad FN(E) = \sum_{(t,F)\in db} |F| - TP(E) \quad FP(E) = \sum_{(t,F)\in db} |E(t)| - TP(E)$$

At the database level, all different instances of an expected fact are counted as one. Then we have instead

$$TP(E) = \left| \bigcup_{(t,F)\in db} E(t) \cap F \right| \quad FN(E) = \left| \bigcup_{(t,F)\in db} F \right| - TP(E) \quad FP(E) = \left| \bigcup_{(t,F)\in db} E(t) \right| - TP(E)$$

However, it is not straightforward to compare two information extractors each characterized by a pair of numbers. The usual mechanism in diagnostic systems is to generate a range of precision numbers over a range of recall numbers to derive a single number called the area under the relative operating characteristic curve (the aROC number) [33] and compare the aROC numbers of two systems. Unfortunately, it is not always possible to obtain aROC numbers of the information extractors we are considering because they are typically not based on a continuous decision threshold. In order to choose an alternative, two conditions should be imposed [34]. The first condition is that it must be able to distinguish the ideal information extractor from the worst information extractor. The second condition is that it shows a gradual and strictly monotonic change in value when the information extractor is changed from the worst to the best one. Note that neither recall nor precision alone satisfies these two conditions.

Many choices that satisfy these two conditions are possible [34, 35]. However, many of them depend on the definition for "true negatives", which is not available in our context. So we propose a variation of the simple matching coefficient (SMC) which simply measures the probability of the information extractor correctly extracting a fact.[2] It is defined as follows and is easily verified to satisfy the two conditions above:

$$SMC(E) = \frac{TP(E)}{TP(E) + FP(E) + FN(E)},\ 0 \le SMC(E)\ \le 1$$

Now we return to our three information extraction tasks. The first task is obvious. We want to recognize proper names of proteins, drugs, and other molecules mentioned in texts. We do not want to recognize names of authors, processes, and any other entities mentioned in these texts.

The second task is slightly more complicated. We want to recognize interaction events between proteins, drugs, or other molecules. These events should include events of transcription, translation, post translational modification, complexing, dissociation, and other similar interactions. If we view each fact as a highly standardized short sentence, we can propose a grammar for them:

PositiveEvent	::=	P phosphorylate P [on T] [at L]
	|	P dephosphorylate P [on T] [at L]
	|	P ubiquinate P
	|	P acetylate P
	|	...
	|	P interact-with P [to-produce P]
	|	P [at L] bind-to P [at L] [to-produce P]
	|	P dissociate [to-produce P+]
	|	P degrade P
	|	P activate-transcription P [to-produce P]
	|	P inhibit-transcription P
	|	P activate [F activity-of] P
	|	P inhibit [F activity-of] P
	|	P transport P [from C] [to C]
Event	::=	PositiveEvent [mediated-by P+] [independent-of P+]
	|	not PositiveEvent [mediated-by P+] [independent-of P+]

In the grammar above, P denotes proteins, drugs, or other molecules; T denotes amino acids; L denotes positions; F denotes biological function; and C denotes cellular locations. In evaluating an information extractor for this task, we can further consider its performance with or without extracting the optional components in the grammar. In the few clauses where a plurality of P's are expected (ie. the $P+$'s), we can consider the situation of a complete or an incomplete match.

[2] A related metric has been proposed in the spoken language processing community for measuring transcription accuracy for transcribing audio input into text (*word error rate*) and for identifying entities and relations among entities (*slot error rate*) [36]. The error rate is the ratio of insertions, deletions and substitution errors divided by the true positives. In our context, we can interpret insertions as false positives and deletions as false negatives; substitutions are not directly relevant. Another related measure is the F-measure, defined as the harmonic mean of recall and precision. That is, $F(E) = (2 \times \mathit{recall}(E) \times \mathit{precision}(E))/(\mathit{recall}(E) + \mathit{precision}(E))$. After substituting the definitions for recall and precision, this reduces to $F(E) = (2 \times TP(E))/(2 \times TP(E) + FN(E) + FP(E))$. There is no intuitive statistical reason for having the multiplicative factor of 2 on $TP(E)$. However, if we drop this multiplicative factor, the result is precisely $SMC(E)$.

It is important to understand that this grammar is not intended as a grammar for parsing scientific texts. Rather, it is more appropriately treated as a grammar defining a set of target representations; the goal is to convert pertinent parts of scientific texts into these relations. As such, an information extractor should convert or normalize different expressions of the same fact into the semantically closest standard form in the grammar. It should not make fine distinctions between different sentence forms. For example, it should convert "camptothecin, an inhibitor of human TOP1" to "camptothecin inhibit TOP1". It should also not make fine distinctions between shades of meanings. For example, "caspase-8 was also stimulated by NB-506" is mapped to "NB-506 activate caspase-8".

The third task is to recognize relationships between the basic events already outlined above. In contrast to the basic events which focus on interactions between molecules, this task is focused on the causal and temporal relationships between two such events. The grammar we propose for them is given below.

Relationship	::=	Event [is-caused-by Event+] [provided Event+]
	|	Event [is-independent-of Event+] [provided Event+]
	|	Event [is-inhibited-by Event+] [provided Event+]

The intention of a relationship such as "E_1 is-caused-by E_2 provided E_3" is as follows. The event E_3 is assumed to have taken place some time in the past and its resultant conditions have remained true. This allows event E_2 to take place and as a result the event E_1 will also take place at the completion of E_2. Again, an information extractor should convert or normalize different expressions of the same event relationships into the semantically closest standard form in the grammar. For example, statement A8 in the appendix of Kohn [37], "c-Abl tyrosine kinase activity is blocked by pRb, which binds to the c-Abl kinase domain", would be mapped according to the grammar as "(pRb inhibit tyrosine kinase activity-of c-Abl) is-caused-by (pRb bind-to c-Abl at kinase domain)."

Having now described the three tasks, we would also like to propose some candidates for forming the benchmark databases for these tasks. We would like to suggest that the appendix of the paper by Kohn [37] as one of the candidates. This appendix lists about 200 statements of interaction events and has sentences of a fairly complex form. Another candidate is the set of MEDLINE abstracts matching the term "Topoisomerase inhibitors." Presently this set includes just over 200 abstracts. A preliminary analysis shows that they contain less than 1000 names and less than 200 interaction events. These numbers are small enough for a small team of experts to construct a benchmark database manually.

6 Automated Database Curation and Ontology Development

Automated database curation and ontology development represents a second challenge application. Automated curation is important because the rate of published experiments is outstripping the ability of database curators to keep up with the relevant literature. In addition, automated curation techniques could allow curators to check the consistency and completeness of their databases. Associated with the curation problem is the database interoperability issue arising from the voluminous, heterogeneous, and distributed data. There is a growing recognition that the adoption of standard nomenclature, controlled vocabulary, and common ontologies are critical to interoperation and integration of biological data. Data integration into a knowledge base system is necessary for answering complex biological questions that may typically involve querying multiple sources. In particular, interesting relationships between database objects, such as relationships among protein sequences, families, structures, and functions, can be readily revealed. An ontology is a semantic model that contains a shared vocabulary and classification of concepts in a domain. An ontology is valuable for query expansion in mining scientific literature and integrating data from heterogeneous sources. It is also important for consistent database curation where functional conservation is annotated with a common language.

Database curation is interesting for another reason: curated databases represent a repository of "gold standard" data. A database entry is typically associated with the literature references from which it is derived—this means that the human curator has already done the extraction from the literature. Craven and Kumlien [38] reported an experiment in which they were able to use the subcellular localization field of the Yeast Protein Database [39]. They collected instances of this relation from the database, traced the references associated with each database entry back to the PubMed abstract, and then within each abstract, identified, where possible, the sentence within the abstract that gave rise to the annotation. This gave them a set of extracted relations (from the database) and the underlying text sources (sentences from the abstracts). They were then able to train and compare several classifiers that extracted the desired localization information. This experiment is suggestive of the ways in which curated databases can be exploited to create "cheap" annotated corpora. It is relatively straightforward to associate the entry in a database field with the underlying article from which it is derived. The harder part is to provide an explicit linkage from the database entry to the phrases and sentences from which it is derived. When the database uses a controlled vocabulary or an ontology to define legal entries for each field, the phrases appearing in the journal article or abstract may not correspond to the actual entry in the database.

In the examples below, we see some of the possible relations between the mention in the literature and its representation in the fields of the database. The example is from the FlyBase genome database [40], where each entry report contains attributed data with hotlinks to the articles from which the information was derived. The first list shows three fields from the polypeptide Appl+P130kD (FBpp0002057) entry, each containing an associated reference [41].

1. Protein size (kD): Luo et al, 1990 130
2. Cell location: Luo et al, 1990 axon
3. Expression pattern: Luo et al, 1990

Stage	Tissue/Position
Embryo	Embryonic Central Nervous System
Embryo	Peripheral Nervous System

The next list contains sentences extracted from the abstract of the Luo et al. [41] article cited above. The phrases in boldface show the specific source of information within each sentence.

1. APPL is synthesized as a ... precursor that is converted to a **130-kDa** secreted for ...
2. APPL immunoreactivity was observed in ... **axonal** tracts, ...
3. In the **embryo**, APPL proteins are expressed exclusively in the **CNS** and **PNS** neurons ...

Even in this very small sample, we see that simple pattern matching suffices in phrase 1 to find *130-kDa*, complex morphology is needed in phrase 2 to associate *axonal tract* with "cell location: axon", and we must decode abbreviations (*CNS* = central nervous system, *PNS* = peripheral nervous system) as well as using information derived from multiple parts of the sentence in phrase 3. A larger sample would contain many more complex mappings between database fields and the underlying literature reference, including entries that require resolution of coreference across sentences or entries that require an analysis of the underlying syntactic relations among entities.

To create an annotated corpus from a curated database, we need to map from entries in database fields back to the text. To do automated database curation, we need the inverse mapping from free text to database entry. We believe that we can create a reversible set of tools that can be used in either direction: mapping from database field to literature or mapping from literature to database. By providing collections of linked pairs of database entry and associated text for use as training and evaluation sets, we would enable many researchers to participate in building tools to automate the database curation process. Although the database curation application is different in its structure from the biological pathway detection experiment outlined in Section 5, it is amenable to the same kinds of automated evaluation techniques outlined there.

As with automated database curation, ontology development can also exploit knowledge accumulated in curated databases and literature. Presented below is a potential knowledge base for protein name ontology. A protein name ontology might be constructed out of a data dictionary and thesaurus of terms and their relationships. Such an ontology is important, because the protein name is the form in which a protein object is referred to and communicated in the scientific literature and biological databases. There is, however, a long-standing problem of nomenclature for proteins, where profligate and undisciplined labeling is hampering communication, as discussed in Nature 1997 [42]. Scientists may name a newly discovered or characterized protein based on its function, sequence features, gene name, cellular location, molecular weight, or other properties, as well as their combinations or abbreviations. Ontology development, nevertheless, requires knowledge acquisition from scientific literature and substantial human effort. Established natural language processing technologies in information extraction, classification and ontology induction can be applied to the protein domain for automated construction of synonym relations among protein names and subsequent classification in terms of the GO functional hierarchy. This would permit greatly enhanced retrieval, using the many synonyms and hypernyms (superordinates) for a given protein name.

A protein knowledge base suitable for ontology development could be developed from curated PIR protein databases (PIR-NREF and iProClass) [44], Gene Ontology (GO) [32], and information extracted from MEDLINE abstracts (Figure 1). The PIR-NREF (Non-redundant REFerence) sequence database contains composite protein names and a bibliography of all published protein sequences. The bibliography is hypertext-linked to PubMed for direct online abstract retrieval. The iProClass (integrated Protein Classification) database [43] provides comprehensive superfamily-domain-motif relationships and structural and functional features of proteins, with links to GO via enzyme (EC) number and keywords. Given the association of database entries and underlying articles, the knowledge base would be ideal for creating annotated corpora containing both protein names (terms) and relationships (isa, homologous_to, has_function) among the protein terms.

The PIR-NREF database provides a timely and comprehensive collection of all protein sequence data, keeping pace with the genome sequencing projects and containing source attribution and minimal redundancy. It currently consists of more than 850,000 sequences from PIR, SwissProt [45], TrEMBL, RefSeq [46], GenPept, and PDB [47], clustered by sequence identity and taxonomy at the species level. The NREF database can be used to develop an initial ontology of protein names and to identify the relationships among terms (e.g., isa, function_of). By looking across annotations for related proteins, it may be possible to detect annotation errors or discrepancies (Figure 2).

The composite protein names from all underlying protein databases, including synonyms, alternate names, and even misspellings, constitute an *initial dictionary* of terms that can help ontology development. As illustrated in Figure 2A, a protein may be variably named based on its function at different hierarchical levels ("ATP-dependent RNA helicase" vs. "RNA helicase"), motif sequence similarity ("DEAD/H box-5"), molecular weight ("protein p68"), or combinations of names ("RNA helicase p68"). A protein may also be named based on gene name. The gene-based names are organism-specific, whereas function- or similarity-based names are usually applied across wide taxonomy ranges. For example, "DRH1" and "DBP2" are names specific to *Arabidopsis thaliana* and *Saccharomyces cerevisiae* (reports not shown), respectively. The different protein names assigned by different source databases may also reveal *relationships among terms* or *annotation errors*. The mixed occurrences of two protein names, "eukaryotic translation initiation factor eIF-4A" and "RNA helicase", (Figure 2B) for the protein entry is a reflection of sequence similarity (i.e., a relationship) shared by common domains and motifs, as described both in iProClass report and MEDLINE abstracts (see below). In another example (Figure 2C), the protein entry is named by three sources as a single-function (EC 3.5.4.19), bifunctional (EC 3.5.4.19, 3.6.1.31), and trifunctional (EC 3.5.4.19, 3.6.1.31, 1.1.1.23) protein. The source name attribution, thus, provides clues for potentially incompletely or mis-annotated proteins.

The composite bibliography information in NREF with PubMed cross-references can be used for direct online abstract retrieval and extraction of additional synonyms or related terms. Through hyperlinks in the databases, one can trace back to the MEDLINE abstracts to identify the sentences within each abstract that contain the protein names and/or their relationships. Information extraction techniques could be used to associate a set of relations (from the database) with their underlying text sources (sentences from the abstracts). Once these are linked, they can be used as training and test sets for developing extraction systems for protein names and relations. These data sets can also be used as test cases for automated classification into existing ontologies (e.g., GO) or even development of ontology induction algorithms.

Figure 3 shows terms and relationships retrievable from MEDLINE abstracts based on curated protein names in the NREF database. The reference (PMID: 2451786) cited in the entry shown in Figure 2A asserts that "p68" has extensive homology with "translation initiation factor eIF-4A" and that "eIF-4A" acts as an "ATP-dependent RNA helicase". Here, the phrase "acts as" implies a new functional relationship between "eIF-4A" and "RNA helicase", extending beyond the extensive homology relationship, and provides a basis for the interchangeable use of the two names in the example entry in Figure 2B.

The GO consists of three ontologies for molecular functions, cellular locations, and biological processes. The terms, together with their definitions and synonyms, are organized in a network architecture (with one or multiple paths to terms). As majority of proteins are named based on their functions or similarity to proteins of known functions, aligning protein names to the widely adopted molecular function GO will help address the database interoperability issue for protein nomenclature. The mapping of protein names to the GO functional hierarchy can also help resolve names that reflect functional characterization at different granularity or alternative functions. For example, "ATP-dependent RNA helicase" is identified as *a kind of* (isa) "RNA helicase" (higher hierarchical level, Figure 4A). As stated in reference PMID: 9592148 (abstract not shown), the protein has also been identified alternatively as a kind of "ATPase" (alternative path to term, Figure 4B).

Data sets for training and validating ontology classification algorithms and ontology induction techniques could be compiled from the protein knowledge base assembled from NREF, using synonyms, protein families and functional relations to improve search accuracy and database curation. As discussed above, composite names in NREF may reveal annotation errors. Thus it is critical to differentiate valid from incorrect names when compiling data sets. Protein family classification allows systematic detection of genome annotation errors when it is based on both global and local similarities at the superfamily (whole protein), domain (structural/functional unit), and motif (structural/functional site) levels. Such comprehensive family relationships are described in iProClass. Figure 5 shows the iProClass report of the "ATP-dependent RNA helicase" NREF entry in Figure 2A. It clearly indicates that the entry is a member of the PIR superfamily [48] SF001321, with several characteristic sequence features, including two Pfam domains [49] (PF00270 and PF00271), one ProClass/ProSite motif [50] (PCM00039), and three sites (nucleotide-binding motifs A and B, and DEAD motif).

In order to benchmark how such extended ontologies could improve retrieval, database interoperability and consistency checking, "gold-standard" data sets could be generated using members of well-characterized protein families that contain positive identifications of sequence features. The retrieval performance using these "gold standard" sets could be compared to retrieval performance of automatically derived synonym sets and ontologies for that set of proteins.

7 Conclusion

This review illustrates both the promise of literature data mining and the need for challenge evaluations. It shows how current language processing approaches can be successfully used to extract and organize information from the literature. It also illustrates the diversity of applications and evaluation metrics. By defining several biologically important challenge problems and by providing the associated infrastructure (annotated data and a common evaluation framework), we can accelerate progress in this field. This will allow us to compare approaches, to scale up the technology to tackle important problems, and to learn what works and what areas still need work.

We should also point out that in this review we have primarily used papers from *Proceedings of Pacific Symposium on Biocomputing* as this has so far been the only conference that has a dedicated track on natural language processing in biology and the papers reported therein were quite representative of the progress made. There are other papers [51, 52, 38, 53, etc.] that we did not discuss and they would be worth further reading to gain a more comprehensive understanding of the field.

References

[1] H. Ogata et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27(1):29–34, January 1999.

[2] I. Xenarios, D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg. DIP: The database of interacting proteins. *Nucleic Acid Res*, 28(1):289–291, January 2000.

[3] G.D. Bader, I. Donaldson, C. Wolting, B.F. Ouellette, T. Pawson, and C.W. Hogue. BIND–the biomolecular interaction network database. *Nucleic Acids Res*, 29(1):242–245, January 2001.

[4] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Proc. Pacific Symposium on Biocomputing'98*, pages 707–718, Maui, Hawaii, January 1998.

[5] B.J. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. In *Proc. Pacific Symposium on Biocomputing*, pages 529–540, 2000.

[6] S.-K. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, 10:104–112, December 1999.

[7] J.C. Park, H.S. Kim, and J.J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *Proc. Pacific Symposium on Biocomputing*, pages 396–407, 2001.

[8] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. In *Proc. Pacific Symposium on Biocomputing*, pages 408–419, 2001.

[9] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. Mining MEDLINE: Abstracts, sentences, or phrases? In *Proc. Pacific Symposium on Biocomputing*, 2002. To appear.

[10] U. Hahn, S. Schulz, and H. Schauer. Rich knowledge capture form medical documents in the MEDSYNDIKATE system. In *Proc. Pacific Symposium on Biocomputing*, 2002. To appear.

[11] G. Leroy and H. Chen. Automated extraction of medical knowledge using underlying logic from medical abstracts. In *Proc. Pacific Symposium on Biocomputing*, 2002. To appear.

[12] J. Pustejovky and J.M. Castano. Robust relational parsing over biomedical literature: Extracting inhibit relations In *Proc. Pacific Symposium on Biocomputing*, 2002. To appear.

[13] B.J. Stapley, L.A. Kelley, M.J.E. Sternberg. Predicting the subcellular location of proteins from text using support vector machines. In *Proc. Pacific Symposium on Biocomputing*, 2002. To appear.

[14] W.J. Wilbur. A thematic analysis of the AIDS literature. In *Proc. Pacific Symposium on Biocomputing*, 2002. To appear.

[15] F. Eisenhaber and P. Bork. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, 15:528–535, 1999.

[16] DARPA (Defense Advanced Research Projects Agency). *Proc. 7th Message Understanding Conference (MUC-7)*, 1998. Available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc.

[17] E. M. Voorhees and D. K. Harman, eds. *Proc. 9th Text REtrieval Conference (TREC-9)*, NIST (National Institute of Standards and Technology) Special Publication 500-XXX, 2001. Available at http://trec.nist.gov/pubs/trec9/t9_proceedings.html

[18] L. Hirschman. The Evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12:281–305, 1998.

[19] DARPA (Defense Advanced Research Projects Agency). *Proc. 6th Message Understanding Conference (MUC-6)*, Columbia, Maryland, Morgan Kaufmann, 1995.

[20] E. M. Voorhees and D. K. Harman, eds. *Proc. 8th Text REtrieval Conference (TREC-8)*, NIST (National Institute of Standards and Technology) Special Publication 500-246, 2000.

[21] E. M. Voorhees and D. M. Tice. The TREC-8 question answering track evaluation. *Proc. 8th Text REtrieval Conference (TREC-8)*, NIST (National Institute of Standards and Technology) Special Publication 500-246, pages 83–105, 2000.

[22] T. Ohta, Y. Tateishi, N. Collier, C. Nobata, and J. Tsujii. Building an annotated corpus from biology research papers. *Proc. COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, Luxembourg, pages 28–34, 2000.

[23] N. Collier, C. Nobata, and J. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. *Proc. 18th International Conference on Computational Linguistic*, Saarbrucken, pages 201–207, 2000.

[24] W.J. Wilbur et al. Analysis of biomedical text for biochemical names: A comparison of three methods. *Proc. AMIA Symposium'99*, pages 176–180, 1999.

[25] L. Wong. PIES, a protein interaction extraction system. In *Proc. Pacific Symposium on Biocomputing*, pages 520–531, 2001.

[26] K. Baclawski, J. Cigna, M.M Kokar, P. Mager, and B. Indurkhya. Knowledge representation and indexing using the unified medical language system. In *Proc. Pacific Symposium on Biocomputing*, pages 493–504, 2000.

[27] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *Proc. Pacific Symposium on Biocomputing*, pages 502–513, 2000.

[28] I.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter. EDGAR: Extraction of drugs, genes, and relations from the biomedical literature. In *Proc. Pacific Symposium on Biocomputing*, pages 514–525, 2000.

[29] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. Automatic extraction of protein interactions from scientific abstracts. In *Proc. Pacific Symposium on Biocomputing*, pages 538–549, 2000.

[30] J.T. Chang, S. Raychaudhuri, and R.B. Altman. Including biological literature improves homology search. In *Proc. Pacific Symposium on Biocomputing*, pages 374–383, 2001.

[31] I. Iliopoulos, A.J. Enright, and C.A. Ouzounis. TEXTQUEST: Document clustering of MEDLINE abstracts for concept discovery in molecular biology. In *Proc. Pacific Symposium on Biocomputing*, pages 384–395, 2001.

[32] M. Ashburner et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.*, 25(1):25–29, 2000.

[33] J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, June 1988.

[34] V. B. Bajic. Comparing the success of different prediction software in sequence analysis: A review. *Briefings in Bioinformatics*, 1(3):214–228, 2000.

[35] M.R. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, 1973.

[36] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance Measures for Information Extraction. *Proc. DARPA Broadcast News Workshop*, pages 249–254, Herndon, VA, Morgan Kaufmann, 1999.

[37] K. W. Kohn. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10:2703–2734, August 1999.

[38] M. Craven and J. Kumlien. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. *Proc. 7th International Conference on Intelligent Systems in Molecular Biology (ISMB-99)*, 1999.

[39] P. E. Hodges, W. E. Payne and J. I. Garrels. Yeast Protein Database (YPD): A database for the complete proteome of the saccharomyces cerevisiae. *Nucleic Acids Research* 26:68–72, 1998.

[40] The Flybase Consortium The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Research* 27:85-88, 1999. Available at http://flybase.bio.indiana.edu/

[41] L.Q. Luo, L. Martin-Morris, and K. White. Identification, secretion, and neural expression of APPL, a Drosophila protein similar to human amyloid protein precursor. *J. Neuroscience* 10(12):3849–3861, 1990.

[42] Obstacles of nomenclature. *Nature* 389:1, 1997.

[43] C.H. Wu, C. Xiao, Z. Hou, H. Huang, and W.C. Barker. iProClass: An integrated, comprehensive, and annotated protein classification database. *Nucleic Acids Research* 29:52–54, 2001.

[44] C.H. Wu, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, R.S. Ledley, K.C. Lewis, H.-W. Mewes, B.C. Orcutt, B.E. Suzek, A. Tsugita, C.R. Vinayaka, L.-S. Yeh, J. Zhang, and W.C. Barker. The Protein Information Resource: An integrated public resource of functional annotation of proteins. *Nucleic Acids Research* 30:35–37, 2002.

[45] A. Bairoch and R. Apweiler. The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28:45–48, 2000.

[46] K.D. Pruitt and D.R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* 29:137–140, 2001.

[47] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research* 28:235–242, 2000.

[48] W.C. Barker, F. Pfeiffer, and D.G. George. Superfamily classification in PIR-International Protein Sequence Database. *Methods in Enzymology* 266:59–71, 1996.

[49] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, E.L.L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Research* 28:263–266, 2000.

[50] K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch. The PROSITE database, its status in 1999. *Nucleic Acids Research* 27:215–219, 1999.

[51] M.A. Andrade and A. Valencia. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics* 14(7):600–607, 1998.

[52] C. Blaschke and M.A. Andrade and C. Ouzounis and A. Valencia. Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Proc. 7th International Conference on Intelligent Systems for Molecular Biology*, pages 60–67, 1999.

[53] E.M. Marcotte and I. Xenarios and D. Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics* 17(4):359–363, 2001.