BIOINFORMATICS REVIEW

Vol. 18 no. 12 2002
Pages 1553–1561

Accomplishments and challenges in literature data mining for biology

Lynette Hirschman[1], Jong C. Park[2], Junichi Tsujii[3], Limsoon Wong[4,*] and Cathy H. Wu[5]

[1]The MITRE Corporation, USA, [2]KAIST, Korea, [3]University of Tokyo, Japan, [4]LIT, Singapore and [5]Georgetown University Medical Center, USA

Received on January 28, 2002; revised on May 20, 2002; accepted on May 23, 2002

ABSTRACT

We review recent results in literature data mining for biology and discuss the need and the steps for a challenge evaluation for this field. 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, and so on. To encourage participation and accelerate progress in this expanding field, we propose creating challenge evaluations, and we describe two specific applications in this context.

Contact: wuc@georgetown.edu; tsujii@is.s.u-tokyo.ac.jp; park@nlp.kaist.ac.kr; lynette@mitre.org; limsoon@lit.org.sg.

INTRODUCTION

Despite the rapid growth in number and size of sequence databases, 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 tables. Today, new kinds of databases that contain information beyond simple sequences are needed, e.g. cellular localization and protein-protein interactions. The forerunners include KEGG (Kanehisa *et al.*, 2002), DIP (Xenarios *et al.*, 2002), and BIND (Bader *et al.*, 2001). They are still small in size and are largely hand curated. The development of good literature data mining technologies can accelerate their growth.

Here, we review literature data mining in biology. The earliest works focused on tasks needing limited linguistic context and processing at the level of words, like identifying protein names (Fukuda *et al.*, 1998) or on tasks relying on word co-occurrence (Stapley and Benoit, 2000) and pattern matching (Ng and Wong, 1999). Later we see linguistic techniques that could handle relations in complex sentences (Park *et al.*, 2001; Yakushiji *et al.*, 2001). Finally we see the emergence of natural language technologies that can handle more complex relations across sentences (Ding *et al.*, 2002; Hahn *et al.*, 2002; Leroy and Chen, 2002; Putejovsky and Castano, 2002; Stapley *et al.*, 2002; Wilbur, 2002).

It is apparent from these papers that there is no common yardstick for assessing and comparing the performance of these systems. Nor are there tools that can be easily applied to biologists' diversed needs for information extraction and text data mining. It is crucial to the development of the field to set up biologically significant challenge problems and corresponding evaluation benchmarks, for both technical component-level and user-centered evaluations. Let us draw from the experience in the newswire domain, where literature data mining techniques have been successful. For example, results from various evaluations show that information extraction systems can identify and classify names of person, organization, location, etc. at accuracies exceeding 90%; and they can successfully extract binary relations among these at over 75% accuracy (DARPA, 1998; Aone *et al.*, 1998).

Much of that progress has arisen due to systematic *common evaluations* conducted at the Message Understanding Conferences and Text REtrieval Conferences (Hirschman, 1998). A 'challenge evaluation' for literature data mining in biology can benefit us, just as the CASP evaluation has accelerated progress in computational protein folding. Here, we also discuss the goals of a challenge evaluation and the ingredients for a successful evaluation. Then we show how a challenge evaluation can be set up in the areas of biological pathway extraction and automated database curation.

RECENT ACCOMPLISHMENTS: NATURAL LANGUAGE PROCESSING PERSPECTIVE

We provide a brief survey on extracting interactions between proteins, drugs, and other molecules. The surveyed works illustrate the progress of the field and show the increasing complexity of the relations that can be extracted.

*To whom correspondence should be addressed.

Fukuda *et al.* (1998) pioneered identification of protein names. They encountered many protein names that were 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. Their solution was to use special properties such as the occurrence of uppercase letters, numerals, and special endings to pinpoint protein names. Similar work at Molecular Connections Pvt Ltd of India suggested that a specificity of over 70% was easily obtainable at an estimated sensitivity of over 70%. The development of a large biology-specific corpus by Ohta *et al.* (2000), and techniques like Hidden Markov Models (Collier *et al.*, 2000) or Bayesian classifiers trained on *k*-grams (Wilbur *et al.*, 1999), would further raise sensitivity and specificity in recognizing protein names.

The field has now progressed beyond recognizing names and has entered the realm of recognizing interactions between proteins and other molecules. The early works could be roughly divided into two main approaches. The first approach, represented by Stapley and Benoit (2000), extracted co-occurrences of gene names from MEDLINE documents and used them to predict their connections based on their occurrence statistics. This approach was followed up by Ding *et al.* (2002), 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 and Wong (1999), used templates that matched specific linguistic structures to recognize and extract protein interaction information from MEDLINE documents.

The work of Ng and Wong (1999) was followed by natural language processing techniques of increasing sophistication. Wong (2001) expanded the number of templates to increase sensitivity. Park *et al.* (2001) introduced a bidirectional incremental parsing technique based on combinatory categorial grammar. Yakushiji *et al.* (2001) used a full parser with a large-scale general-purpose grammar to analyze MEDLINE abstracts. While no extensive validation results were available on these systems, their specificity was estimated at 60–80%. Another recent system is GENIES (Friedman *et al.*, 2001). It extracts a broad set of biological relations, including embedded relations. It reached 96% precision at 63% recall on a hand-annotated 8 000 word article from *Cell*. All these papers could handle sentences whose structures were more complex than those of Ng and Wong (1999). However, none of them handle pronouns. Their overall sensitivity also remains an issue.

Most recently, research has gone beyond treatment of single sentences to look at relations that span multiple sentences through the use of coreference. Putejovsky and Castano (2002) focused on relations of the word *inhibit* and showed that it was possible to extract biologically important information from free text reliably, using a corpus-based approach to develop rules specific to a class of predicates. A strength of this system was its anaphora resolution module. Hahn *et al.* (2002) described the MEDSYNDIKATE system for acquiring knowledge from medical reports. It could analyze co-referring sentences and extract new concepts given a set of grammatical constructs. Leroy and Chen (2002) presented GeneScene. It used prepositions as entry points into phrases in the text, in contrast to the main trend that used verbs. It then filled in a set of templates of patterns of prepositions around verbs and nominalized verbs. It also had rules for combining these templates to extract information from more complex sentences. Based on small-scale experiments, these systems should have higher performance. For example, Putejovsky and Castano (2002) reported 90% specificity at 57% sensitivity for extraction of 'inhibit' relations.

However, it is unclear how to compare the different approaches; it is also unclear how well a system has to perform to be useful. To compare technical approaches, different systems must be applied to the same domain via common evaluations. To know how good a system has to be, prototypes must be given to biologists in user-centered evaluations. As learned from previous evaluations in the information retrieval community (Hersh *et al.*, 2001), it is hard to extrapolate from results of batch experiments to predict complex issues of utility and user acceptance of interactive tools. However, even imperfect tools are useful, if they give improved functionality at low cost.

RECENT ACCOMPLISHMENTS: BIOMEDICAL APPLICATIONS

Besides the recognition of protein interactions from scientific text, natural language processing has been applied to a broad range of information extraction problems in biology. We briefly describe here some of these results.

We begin with systems that capture specific relations in databases. Hahn *et al.* (2002) used natural language techniques and nomenclatures of the Unified Medical Language System (UMLS) to learn ontological relations for a medical domain. Baclawski *et al.* (2000) is a diagrammatic knowledge representation method called keynets. The UMLS ontology was used to build keynets. Using both domain-independent and domain-specific knowledge, keynets parsed texts and resolved references to build relationships between entities. Humphreys *et al.* (2000) described two information extraction applications in biology based on templates: EMPathIE extracted from journal articles details of enzyme and metabolic pathways; PASTA extracted the roles of amino acids and active sites in protein molecules. This work illustrated the importance of template matching, and applied the tech-

nique to terminology recognition. Rindflesch *et al.* (2000) described EDGAR, a system that extracted relationships between cancer-related drugs and genes from biomedical literature. EDGAR drew on a stochastic part of speech tagger, a syntactic parser able to produce partial parses, a rule-based system, and semantic information from the UMLS. The metathesaurus and lexicon in the knowledge base were used to identify the structure of noun phrases in MEDLINE texts. Thomas *et al.* (2000) customized an 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 interactions.

The next papers focus on improving retrieval and clustering in searching large collections. Chang *et al.* (2001) modified PSI-BLAST to use literature similarity in each iteration of its search. They showed that supplementing sequence similarity with information from biomedical literature search could increase the accuracy of homology search result. Illiopoulos *et al.* (2001) gave a method for clustering MEDLINE abstracts based on a statistical treatment of terms, together with stemming, a 'go-list', and unsupervised machine learning. Despite the minimal semantic analysis, clusters built here gave a shallow description of the documents and supported concept discovery. Wilbur (2002) formalized the idea of a 'theme' in a set of documents as a subset of the documents and a subset of the indexing terms so 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.

Finally, text processing has been used for classification. Stapley *et al.* (2002) used a support vector machine to classify terms derived by standard term weighting techniques to predict the cellular location of proteins from description in abstracts. The accuracy 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 a hand-crafted rule-based classifier (Eisenhaber and Bork, 1999).

ORGANIZING A CHALLENGE EVALUATION

We showed earlier the potential of literature data mining techniques in biology. However, few of these techniques have made it into routine use to help manage biological information. To know which techniques will work on which problems, we need a systematic evaluation. We can do this via biologically motivated common challenge problems that will attract researchers. A challenge problem would focus on a task 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 to understand what worked and what didn't work. This approach has already produced great progress in the realms of text processing, machine learning, etc. and in biology, protein structure prediction.

We identify the ingredients for a successful evaluation.

CHALLENGE PROBLEM. It must be a problem of biological significance, such as literature search to assemble biological pathways, or creation of specialized databases to organize information.

TASK DEFINITION. This defines the criteria for evaluation—what constitutes a 'correct' answer in the context of the challenge problem, including a formal specification of the target output.

TRAINING DATA. To build systems to solve the 'challenge problem,' developers need annotated data for 'practice tests'—with the right answers provided. The training data could also 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, expressed as *annotations*, would facilitate the creation of information extraction rules that map from the free text occurrence of information to the required target output.

TEST DATA. Once a system is built, it must be evaluated on *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.

EVALUATION METHODOLOGY. 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. In addition, the evaluation methods must also be accompanied by statistical tests to assess the significance of statistical differences among systems.

EVALUATOR. There must be a neutral group who 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, especially the designated evaluator group. Researchers are also more likely to join if

funding is associated with the evaluation, even if the funding is indirect such as a government or private funded program that rewards good results in the evaluation.

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

EXTRACTION OF BIOLOGICAL PATHWAYS

We consider biological pathways as a network of interactions and events between proteins, drugs, and other molecules. We propose three layers of challenges. First, the task is to recognize names of proteins, drugs, and other molecules. Next, the task is to recognize basic interaction events between molecules. Last, the task is to recognize the relationships between the basic interaction events.

Let us first set up the benchmarking framework. It is oriented towards information extraction rather than natural language understanding—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. First, filling in a template is closer to the application scenario of filling in a database table. Second, information extraction need not be syntax based, so this choice allows us to assess a broader range of techniques. Third, the articles may not be written in grammatical English.

The framework is as follows. A number of test databases are built. Each database is a set of records. Each record has a text to be tested and a list of expected facts. The text can be a sentence, an abstract, or a whole article. The list of expected facts are all the correct or actual facts that a 'perfect' information extractor for the task on hand can extract from the given text and nothing else. 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 $\{(t_1, F_1), \ldots, (t_m, F_m)\}$, where t_i are the texts and $F_i = \{f_{i,1}, \ldots, f_{i,n_i}\}$ are expected facts. There are two levels of evaluation: the level of individual records, and the level of the entire test database.

In a traditional evaluation of information retrieval systems, 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

$$recall(E) = TP(E)/[TP(E) + FN(E)]$$
$$precision(E) = TP(E)/[TP(E) + FP(E)].$$

The definitions for $TP(E)$ (true positives), $FN(E)$ (false negatives), and $FP(E)$ (false positives) depend on whether we are evaluating at the record level or at the database level.[†] 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

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

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

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

It is hard to compare two information extractors each characterized by two numbers. The usual method in diagnostic systems is to obtain a range of precision values over a range of recall values to get the area under the relative operating characteristic curve (aROC) and compare the aROC of two systems (Swets, 1988). But it is hard to get the aROC of information extractors we are considering as they often do not have adjustable decision thresholds. Two conditions are imposed in choosing an alternative (Bajic, 2000): to distinguish the ideal information extractor from the worst one, and to show a gradual monotonic change in value when the information extractor is changed from the worst to the best one.

Many choices satisfy these two conditions (Bajic, 2000). However, many of them rely on the definition for 'true negatives' unavailable in our context. So we propose a variation of the simple matching coefficient (SMC).[‡] It is defined below and satisfies the two conditions above:

$$SMC(E) = TP(E)/[TP(E) + FN(E) + FP(E)]$$

[†] It is impossible to define the usual notion of true negatives because there is no theoretical bound on the number of 'facts' that can be generated from a sentence and it is unreasonable to use the closed world assumption here.

[‡] A related metric been proposed in the spoken language processing for measuring transcription accuracy for transcribing audio input into text (*word error rate*) and for identifying entities and relations among entities (*slot error rate*) (Makhoul *et al.*, 1999). The error rate is the sum 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, which is the harmonic mean of recall and precision, $F(E) = (2 \times recall(E) \times precision(E))/(recall(E) + precision(E))$. Applying 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 $SMC(E)$.

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

For the second task, we want to recognize interaction events between proteins, drugs, or other molecules. These events include transcription, translation, post translational modification, complexing, dissociation, etc. If we view each fact as a highly standardized sentence, we can propose a grammar for them below,[§] where *P* stands for proteins, or other molecules; *T* for amino acids; *L* for positions; *F* for biological function; and *C* for cellular locations:

```
PosEvent  ::=  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     ::=  PosEvent [mediated-by P+] [independent-of P+]
           |   not PosEvent [mediated-by P+] [independent-of P+]
```

The grammar is not for parsing. It is a grammar defining a set of target representations; the goal is to convert pertinent parts of scientific texts into these relations. An information extractor should convert 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 TOP1' to 'camptothecin inhibit TOP1'. It should not make fine distinctions between shades of meanings. For example, 'caspase8 was stimulated by NB506' is mapped to 'NB506 activate caspase8'.

The third task is to recognize relationships between basic events. In contrast to basic events which focus on interactions between molecules, this task is focused on the causality between two such events. The grammar we propose for them is:

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

[§] A biological pathway may contain events, such as the opening of a vesicle, that are not molecular interactions. However, it is better to start with a more focused class of events. We choose molecular interactions based on a straw poll of several researchers from the pharmaceutical and biotechnology industry and because these interactions are already familiar from works reported at the past Pacific Symposiums on Biocomputing.

The intention of a relationship like '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 ago and its resultant conditions have remained true. This allows E_2 to take place and thus E_1 will take place at the end of E_2. An information extractor should convert different expressions of the same event relationships into the semantically closest standard form in the grammar. For example, statement A8 in Kohn (1999), 'c-Abl tyrosine kinase activity is blocked by pRb, which binds to the c-Abl kinase domain', would be mapped as '(pRb inhibit tyrosine kinase activity-of c-Abl) is-caused-by (pRb bind-to c-Abl at kinase domain).'

Having described the three tasks, we now propose some candidates for the benchmark databases for these tasks. We suggest the appendix of Kohn (1999) as a candidate. It lists about 200 statements of interaction events and has sentences of a fairly complex form. Another candidate is the set of MEDLINE abstracts on 'Topoisomerase inhibitors.' 150–200 new abstracts on this topic appear in MEDLINE each year. A rough analysis shows that each year's worth of abstracts have less than 1000 names and less than 200 interaction events, small enough for a small team of experts to build a benchmark database manually.

AUTOMATED DATABASE CURATION AND ONTOLOGY DEVELOPMENT

As with automated database curation, ontology development can exploit the knowledge accumulated in curated databases and literature. A protein name ontology may be constructed from a data dictionary and thesaurus of terms and their relationships. Such an ontology is important because the protein name is often how a protein object is referred to 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). Scientists may name a newly discovered 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, therefore, requires knowledge acquisition from scientific literature and substantial human effort. 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 functional hierarchy of Gene Ontology (GO) (Gene Ontology Consortium, 2001). This would permit greatly enhanced retrieval, using the many synonyms and hypernyms (superordinates) for a given protein name.

Database curation is also interesting since curated

databases represent a repository of 'gold standard' data. Craven and Kumlien (1999) reported an experiment where they used the subcellular localization field of the Yeast Protein Database (Hodges *et al.*, 1998). 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 the sentence 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 suggests how curated databases might be exploited to create 'cheap' annotated corpora. It is easy to associate an entry in a database field with the underlying article from which it is derived. It is harder 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 article may not correspond to the actual entry in the database.

We see below some possible relations between the mention in the literature and its representation in the fields of the database. The example is from FlyBase (Flybase Consortium, 2002), where each entry contains attributed data with links to the source articles. The first list shows three fields from the Appl+P130kD (FBpp0002057) entry, each having an associated reference 'Luo et al., 1990'.

(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 Luo *et al.* (1990). The phrases in boldface pinpoint the source of information within each sentence.

(1) APPL ... is converted to a **130-kDa** secreted form ...

(2) APPL ... was observed in ... **axonal** tracts, ...

(3) In the **embryo**, APPL proteins are expressed exclusively in the **CNS** and **PNS** neurons ...

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'. But we must decode abbreviations (*CNS* = central nervous system, *PNS* = peripheral nervous system) and also use information derived from multiple parts of the sentence in phrase 3. A larger sample would have 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.

This exploration has led to a dataset for the Knowledge Discovery and Data Mining Challenge Cup 2002; see http://www.biostat.wisc.edu/~craven/kddcup/tasks.html. The challenge task, based on activities performed by Flybase curators, requires that participants build systems to automatically process the journal articles to answer the following questions, given a collection of articles, each labeled with the genes mentioned in the article:

- Does the article contain any experimental results about gene expression that should be put in the database?
- If so, for each gene in the article, is there experimental evidence for any transcripts (RNA), protein, or polypeptide products of that gene?

The participants are given an 800-document training set of full text articles with their associated Flybase entries. The test set consists of several hundred more articles whose database entries have not yet been published in Flybase, so that they constitute blind test data. The provision of database entries and associated full text for use as training and evaluation sets enables many researchers to participate in building tools for database curation. Each participating system will be evaluated by how well it can distinguish articles that should be curated from those without experimental evidence (a classification task). It will also be evaluated on how well it can determine which *genes* in a given article have information that should be curated, which will require more fine-grained analysis.

For ontology development and challenge tasks, a protein knowledge base could be built from PIR protein databases (iProClass and PIR-NREF) (Wu *et al.*, 2002), GO, and MEDLINE abstracts (Fig. 1[¶]). NREF provides comprehensive protein sequence data with source attribution and minimal redundancy (Fig. 2). It currently contains more than 930 000 sequences organized from PIR, Swiss-Prot (Bairoch and Apweiler, 2000), TrEMBL, RefSeq (Pruitt and Maglott, 2001), GenPept, and PDB (Westbrook *et al.*, 2002). The 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. The iProClass database (Wu *et al.*, 2001) provides comprehensive protein family relationships and structural and functional features, 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

[¶] See http://pir.georgetown.edu/pirwww/doc/bioinf02_figure.pdf for all the figures.

annotated corpora containing both protein names (terms) and relationships (isa, homologous-to, has-function) among the protein terms.

As shown in Fig. 2A, a protein may be named based on its function at different levels ('ATP-dependent RNA helicase' versus 'RNA helicase'), motif sequence similarity ('DEAD/H box-5'), molecular weight ('protein p68'), or combinations of names ('RNA helicase p68'). The different protein names assigned by different source databases may also reveal relationships among terms or annotation errors. The mixed occurrence of two protein names, 'eukaryotic translation initiation factor eIF-4A' and 'RNA helicase' (Fig. 2B) for the protein entry reveals sequence similarity (i.e. a relationship) shared by common domains and motifs. There are also many examples of discrepant annotations (not shown) that provide clues for potentially incompletely or mis-annotated proteins.

The NREF bibliography and associated PubMed links allow online abstract retrieval and extraction of synonyms or related terms by identifying the sentences within abstracts that contain the protein names and their relationships. Information extraction techniques could associate a set of relations (from the database) with their underlying text sources (sentences from the abstracts). The data can be used for developing extraction systems for protein names and relations, for automated classification into existing ontologies (For example, GO), or for development of ontology induction algorithms. Fig. 3 shows terms and relationships retrievable from MEDLINE abstracts based on protein names in NREF. The reference (PMID: 2451786) cited in the entry shown in Fig. 2A asserts that 'p68' has extensive homology with 'translation initiation factor eIF-4A' and that 'eIF-4A' acts as an 'ATP-dependent RNA helicase.' 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 Fig. 2B.

The GO consists of ontologies for molecular functions, cellular locations, and biological processes. The terms are organized in a network. As many proteins are variably named based on their functions or similarity to proteins of known functions, aligning protein names to the widely used 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' (Fig. 4A). As stated in reference PMID: 9592148 (abstract not shown), the protein has also been identified alternatively as a kind of 'ATPase' (Fig. 4B).

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. Protein family classification allows systematic detection of annotation errors when it is based on both global and local similarities at the superfamily (whole protein), domain, and motif levels. Such comprehensive family relationships are described in iProClass. As shown in the iProClass report (Fig. 5), the protein in Fig. 2A is a member of the PIR superfamily SF001321, with several characteristic sequence features, including two domains (PF00270 and PF00271), one motif (PCM00039), and three sites (nucleotide-binding motifs A and B, and DEAD motif).

The evaluation of ontologies is a challenging task, in part because there is no established metric for measuring knowledge in terms of content or value. The protein ontology can be evaluated at two levels (Sparck-Jones and Galliers, 1996). An intrinsic evaluation, where the protein name ontology induction procedure is evaluated without reference to a particular task, may involve the comparisons of the terms and ontological relations discovered by the system against those found by humans. As illustrated in a MITRE prototype (Mani *et al.*, 2002), direct comparison can also be made between a sample of a machine ontology and of a human ontology built for the same domain corpus, with time measurements for the creation of each. An extrinsic (task-oriented) evaluation may evaluate the ontology's usefulness in manual query expansion. Users of the protein-name based bibliography search using PubMed can be offered related terms from the protein name ontology for manual query expansion. The accuracy of retrieval can then be measured by the percentage of relevant documents in the top *n* hits under different types of query expansion.

CONCLUSION

This review shows 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, 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 been the only

conference that has a dedicated track on natural language processing in biology. There are other papers (Andrade and Valencia, 1998; Blaschke *et al.*, 1999; Craven and Kumlien, 1999; Marcotte *et al.*, 2001, etc.) that we did not discuss and they would be worth further reading to gain a more comprehensive understanding of the field.

REFERENCES

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

Aone,C. *et al.* (1998) SRA: Description of the IE2 system used for MUC-7. In *Proc. 7th Message Understanding Conf.*

Baclawski,K. *et al.* (2000) Knowledge representation and indexing using the unified medical language system. *PSB 2000*, 493–504.

Bader,G. *et al.* (2001) BIND–the biomolecular interaction network database. *NAR*, **29**, 242–245.

Bairoch,A. and Apweiler,R. (2000) The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. *NAR*, **28**, 45–48.

Bajic,V.B. (2000) Comparing the success of different prediction software in sequence analysis: A review. *Brief Bioinform.*, **1**, 214–228.

Blaschke,C. *et al.* (1999) Automatic extraction of biological information from scientific text: Protein-protein interactions. *ISMB*, **7**, 60–67.

Chang,J. *et al.* (2001) Including biological literature improves homology search. *PSB 2001*, 374–383.

Collier,N. *et al.* (2000) Extracting the names of genes and gene products with a hidden Markov model. *Int. Conf. Comput. Linguistics*, **18**, 201–207.

Craven,M. and Kumlien,J. (1999) Constructing biological knowledge bases by extracting information from text sources. *ISMB*, **7**, 77–86.

DARPA, (1998) *Proc. 7th Message Understanding Conf.*

Ding,J. *et al.* (2002) Mining MEDLINE: Abstracts, sentences, or phrases? *PSB 2002*, 326–337.

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

Flybase Consortium, (2002) The FlyBase database of the Drosophila genome projects and community literature. *NAR*, **30**, 106–108.

Friedman,C. *et al.* (2001) GENIES: A natural language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**, S74–S82.

Fukuda,K. *et al.* (1998) Toward information extraction: Identifying protein names from biological papers. *PSB 1998*, 707–718.

Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. *Genome Res.*, **11**, 1425–1433.

Hahn,U. *et al.* (2002) Rich knowledge capture from medical documents in the MEDSYNDIKATE system. *PSB 2002*, 338–349.

Harabagiu,S. *et al.* (2001) FALCON: Boosting knowledge for answer engines. In *Proc. 9th Text Retrieval Conf.*

Hersh,W. *et al.* (2001) Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing and Management*, **37**, 383–402.

Hirschman,L. (1998) The evolution of evaluation: Lessions from the Message Understanding Conferences. *Computer Speech and Language*, **12**, 281–305.

Hodges,P. *et al.* (1998) Yeast Protein Database (YPD): A database for the complete genome of the *saccharomyces cerevisiae*. *NAR*, **26**, 68–72.

Humphreys,K. *et al.* (2000) Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. *PSB 2000*, 502–513.

Illiopoulos,I. *et al.* (2001) TEXTQUEST: Document clustering of MEDLINE abstracts for concept discovery in molecular biology. *PSB 2001*, 384–395.

Kanehisa,M. *et al.* (2002) The KEGG database at GenomeNet. *NAR*, **30**, 42–46.

Kohn,K.W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell*, **10**, 2703–2734.

Leroy,G. and Chen,H. (2002) Automated extraction of medical knowledge using underlying logic from medical abstracts. *PSB 2002*, 350–361.

Luo,L. *et al.* (1990) Identification, secretion, and neural expression of APPL, a *Drosophila* protein similar to human amyloid protein precursor. *J. Neuroscience*, **10**, 3849–3861.

Makhoul,J. *et al.* (1999) Performance measures for information extraction. In *Proc. DARPA Broadcast News Workshop*. pp. 249–254.

Mani,I. *et al.* (2002) Automatically inducing ontologies from corpora. Technical note, MITRE.

Marcotte,E.M. *et al.* (2001) Mining literature for protein-protein interactions. *Bioinformatics*, **17**, 359–363.

Nature (1997) Obstacles of nomenclature. *Nature*, **389**, 1.

Ng,S.-K. and Wong,M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *GIW*, **10**, 104–112.

Ohta,T. *et al.* (2000) Building an annotated corpus from biology research papers. In *Proc. COLING-2000 Workshop on Semantic Annotation and Intelligent Content*. pp. 28–34.

Park,J. *et al.* (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *PSB 2001*, 396–407.

Pruitt,K. and Maglott,D. (2001) refSeq and LocusLink: NCBI gene-centered resources. *NAR*, **29**, 137–140.

Putejovsky,J. and Castano,J. (2002) Robust relational parsing over biomedical literature: Extracting inhibit relations. *PSB 2002*, 362–373.

Rindflesch,T. *et al.* (2000) EDGAR: Extraction of drugs, genes, and relations from biomedical literature. *PSB 2000*, 517–528.

Sparck-Jones,K. and Galliers,J. (1996) *LNAI 1083: Evaluating National Language Processing Systems—An Analysis and Review*. Springer.

Stapley,B. and Benoit,G. (2000) Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline asbtracts. *PSB 2000*, 529–540.

Stapley,B. *et al.* (2002) Predicting the subcellular location of proteins from text using support vector machines. *PSB 2002*, 374–385.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Thomas,J. *et al.* (2000) Automatic extraction of protein interactions from scientic abstracts. *PSB 2000*, 538–549.

Westbrook,J. *et al.* (2002) The Protein Data Bank: Unifying the archive. *NAR*, **30**, 245–248.

Wilbur,W. (2002) A thematic analysis of the AIDS literature. *PSB 2002*, 386–397.

Wilbur,W. *et al.* (1999) Analysis of biomedical text for biochemical names: A comparison of three methods. *AMIA Symp 1999*, 176–180.

Wong,L. (2001) PIES, a protein interaction extraction system. *PSB 2001*, 520–531.

Wu,C. *et al.* (2002) The Protein Information Resource: An integrated public resource of functional annotation of proteins. *NAR*, **30**, 35–37.

Wu,C. *et al.* (2001) iProClass: An integrated, comprehensive, and annotated protein classification database. *NAR*, **29**, 52–54.

Xenarios,I. *et al.* (2002) DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *NAR*, **30**, 303–305.

Yakushiji,A. *et al.* (2001) Event extraction from biomedical papers using a full parser. *PSB 2001*, 408–419.