Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry
Computational Biology and Chemistry 28 (2004) 409–416
www.elsevier.com/locate/compbiolchem

Database Note

iProLINK: an integrated protein resource for literature mining

Zhang-Zhi Hu[a], Inderjeet Mani[b], Vincent Hermoso[a], Hongfang Liu[c], Cathy H. Wu[a,*]

[a] *Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, DC 20057, USA*
[b] *Georgetown University, 37th and O Streets, NW, Washington, DC 20057, USA*
[c] *University of Maryland at Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA*

Received 30 September 2004; accepted 30 September 2004

Abstract

The exponential growth of large-scale molecular sequence data and of the PubMed scientific literature has prompted active research in biological literature mining and information extraction to facilitate genome/proteome annotation and improve the quality of biological databases. Motivated by the promise of text mining methodologies, but at the same time, the lack of adequate curated data for training and benchmarking, the Protein Information Resource (PIR) has developed a resource for protein literature mining—iProLINK (integrated Protein Literature INformation and Knowledge). As PIR focuses its effort on the curation of the UniProt protein sequence database, the goal of iProLINK is to provide curated data sources that can be utilized for text mining research in the areas of bibliography mapping, annotation extraction, protein named entity recognition, and protein ontology development. The data sources for bibliography mapping and annotation extraction include mapped citations (PubMed ID to protein entry and feature line mapping) and annotation-tagged literature corpora. The latter includes several hundred abstracts and full-text articles tagged with experimentally validated post-translational modifications (PTMs) annotated in the PIR protein sequence database. The data sources for entity recognition and ontology development include a protein name dictionary, word token dictionaries, protein name-tagged literature corpora along with tagging guidelines, as well as a protein ontology based on PIRSF protein family names. iProLINK is freely accessible at http://pir.georgetown.edu/iprolink, with hypertext links for all downloadable files.

Keywords: PubMed; UniProt; Literature mining; Natural language processing; Post-translation modifications; Protein annotation

1. Introduction

Increasingly researchers have studied complex biological systems on global scales ranging from genomes and proteomes to metabalomes. Full exploration of these valuable data requires advanced bioinformatics infrastructures for biological knowledge management. In particular, major curated databases, such as the UniProt protein knowledgebase (Apweiler et al., 2004) and various genome databases, represent basic resources for biological interpretation of large-scale data. Of special value in these databases are annotations derived from experimentally verified published data that represent the latest scientific knowledge about specific genes and proteins. However, the amount of such literature-based and manually-curated annotation is rather limited due to the laborious nature of knowledge extraction from the literature.

With an ever-increasing volume of scientific literature now available electronically, there is both a pressing need and a great opportunity in developing more efficient ways of literature data mining. Indeed, in recent years, natural language processing (NLP) technologies are being utilized for biological literature mining and information extraction (Hirschman et al., 2002a). As a member of the UniProt consortium, our group at the Protein Information Resource (PIR) (Wu et al., 2003a) is primarily interested in the "database curation" application—namely, extracting experimental information from the scientific literature and populating the data in appropriate annotation fields of the UniProt database.

The process of applying literature mining methods for protein database curation involves several tasks (Fig. 1):

* Corresponding author. Tel.: +1 202 687 1039; fax: +1 202 687 1662.
E-mail address: wuc@georgetown.edu (C.H. Wu).

1476-9271/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2004.09.010

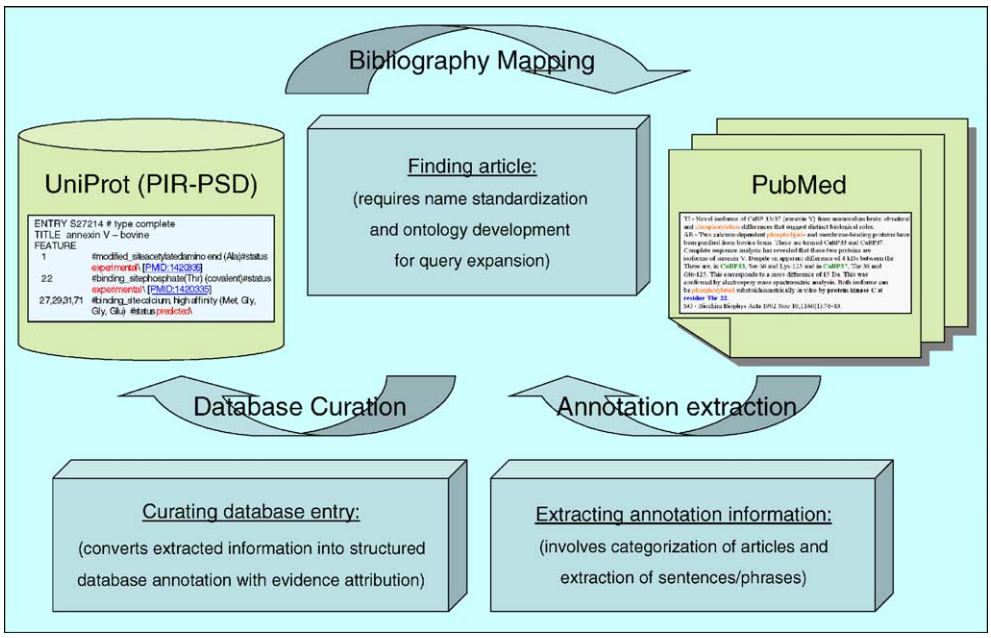

Fig. 1. Bibliography mapping and annotation extraction for literature-based database curation.

- bibliography mapping: identification of articles from literature sources (such as PubMed) that describe a given protein entry;
- annotation extraction: categorization of annotation types and extraction of sentences and/or phrases describing the given annotation; and
- database curation: conversion of the extracted literature information into annotation in the database with structured syntax, controlled vocabulary, and evidence attribution.

These tasks are also related to the topics of protein named entity recognition and protein ontology development. A prerequisite to bibliography mapping is protein named entity recognition–identification of protein names from articles. Furthermore, due to the long-standing problem of protein nomenclature, a protein ontology can assist entity recognition with the description of names and synonyms of protein classes as well as their relationships.

Future progress in biological literature mining and annotation extraction requires close collaboration of computational and biological scientists. Benchmarking data and resources need to be developed for training and evaluating literature mining methodologies, while biological domain experts need to provide scientific validation and explanation for literature mining results. To evaluate the utility of text mining techniques for mining biological data from literature, there have been community evaluation contests such as Knowledge Discovery and Data Mining cup (KDD) (Yeh et al., 2003) and more recently Critical Assessment of Information Extraction systems in Biology (BioCreAtIve) (http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html).

Inspired by the promise of text mining methodologies for database curation, but at the same time, the lack of adequate curated data for training and benchmarking, PIR has developed a resource for protein literature mining—iProLINK (integrated Protein Literature, INformation and Knowledge). This paper describes the various data sources in iProLINK and their application to literature mining research.

2. iProLINK overview

The data sources in iProLINK are organized into two major categories based on their utilization for text mining/NLP research (Fig. 2), as summarized below and detailed in Sections 3 and 4.

2.1. Literature-based protein curation

NLP research for literature-based protein curation involves bibliography mapping (to map literature to protein database entries) and annotation extraction (to extract annotation information from the literature). The corresponding data sources include:

- bibliography system: bibliography information pages for all protein entries with protein to PubMed ID (PMID) mapping, as well as a bibliography submission page for researchers to map and submit papers; and
- literature corpus: abstracts and full-text articles manually tagged with experimentally validated protein features such as post-translation modifications.

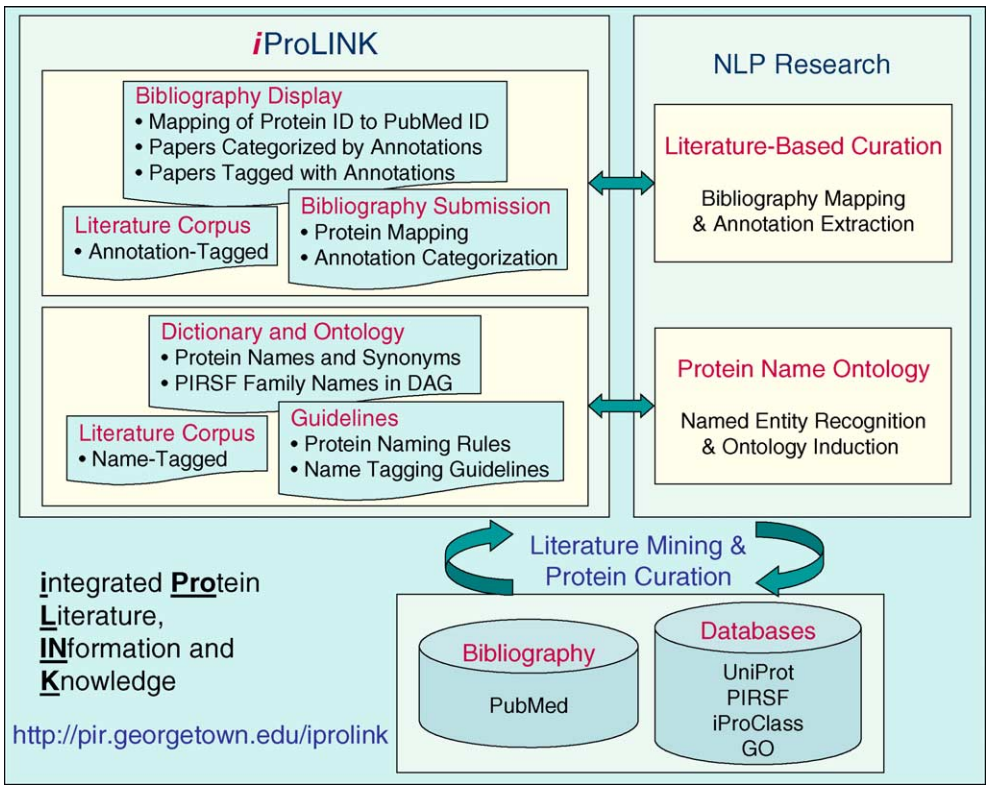

Fig. 2. iProLINK as a resource for text mining research to facilitate protein database curation.

2.2. Protein named entity recognition and ontology development

To facilitate NLP approaches to named entity recognition and ontology development, iProLINK includes:

- protein dictionaries: a protein name dictionary consisting of terms for names, synonyms and acronyms, and word token dictionaries consisting of biomedical terms, chemical terms, macromolecules, common English, and non-word tokens;
- protein ontology: an ontology based on PIRSF (Wu et al., 2004a) protein family names;
- protein and family naming guidelines: documents outlining rules and conventions for assigning protein names and protein family names;
- protein tagging guidelines and Literature Corpus: guidelines for tagging protein names in abstracts, and abstracts manually tagged with protein names.

The resource is freely accessible from the PIR web site at http://pir.georgetown.edu/iprolink, with a listing and hypertext links for all downloadable files. The site also provides search mechanisms for accessing the PIR bibliography system. Protein feature-based searches provide access to papers with tagged experimental feature evidence. Also linked are PIR collaborators who are conducting related text mining/NLP research projects using the iProLINK resource.

3. iProLINK resource for bibliography mapping and annotation extraction

The inclusion of experimentally validated annotation with literature citation for evidence attribution can greatly enhance the quality and value of protein databases. As the volume of sequence data and scientific literature continues to grow exponentially, the manual processes by which the evidence attribution has been done in the past become a bottleneck in protein database curation. It is essential to develop computational approaches for the mapping and extraction of protein experimental data. Indeed, curated databases and their associated PubMed abstracts have been used to create annotated corpora for training classifiers to extract protein localization information (Craven and Kumlien, 1999).

3.1. PIR bibliography system

Linking protein entries to relevant scientific literature that describes or characterizes the proteins is crucial for increasing the amount of experimentally verified data and for improving the quality of protein annotation. To provide a more comprehensive bibliographic coverage for all UniProt protein entries, PIR has developed a bibliography information system. The bibliography system includes a biweekly-updated bibliography database, as well as a web interface for browsing and searching bibliography information pages and for submitting bibliographic data for UniProt proteins.

The bibliography information page provides, for each protein entry, reference citations compiled from multiple sources, including several curated databases. In addition to the underlying UniProt database (with references combined from Swiss-Prot, TrEMBL, and PIR-PSD), bibliographic data are also collected from databases such as SGD (Christie et al., 2004), MGD (Blake et al., 2003) and GeneRIF (Mitchell et al., 2003). Many reference citations are curated, providing categorization of protein annotation information contained in the citation. The bibliography submission interface guides users through steps in mapping reference citations to protein entries, entering the bibliographic data, and summarizing the contents using categories (such as genetics, tissue/cellular localization, molecular complex or interaction, function, regulation, and disease), with evidence attribution (experimental or predicted) and description of methods.

3.2. PIR feature evidence attribution—citation mapping and evidence tagging

In the PIR-PSD database, feature annotations such as binding sites, catalytic sites, and modified sites, are labeled with status tags "*experimental*" or "*predicted*" to distinguish experimentally verified from computationally predicted data. However, such "*experimental*" tags were not originally attributed with literature citations for the experimental evidence, even though the relevant citations are usually present in the Reference section of the PSD sequence report. To appropriately attribute bibliographic data to features with experimental evidence, we have been conducting a retrospective literature survey (Wu et al., 2003b). The evidence-attributed PSD experimental feature data are being incorporated into the UniProt knowledgebase.

The retrospective literature survey involves both citation mapping (finding citations from the Reference section that describe the given experimental feature) and evidence tagging (tagging the sentences providing experimental evidence in an abstract and/or full-text article). Fig. 3 shows an example of evidence attribution for the feature "binding_site: phosphate (Thr) (covalent)" at amino acid residue 22. The attribution includes a direct PMID citation on the feature line (Fig. 3A) and an annotation-tagged text where a sentence is highlighted in the abstract and another sentence is quoted from the full-text paper (Fig. 3B). There are a total of 5296 PSD protein sequence entries with 9788 experimental feature lines to be mapped. Currently, over 3700 feature lines have been manually attributed with literature citation, half of which also have evidence tagging in abstracts and/or full-text articles. The status of citation mapping and evidence tagging for different feature types can be viewed at http://pir.georgetown.edu/cgi-bin/ipkLitFt.pl?stat=1.

The mapped citations and annotation-tagged texts not only provide users with quality annotation, but can also serve as NLP training data. For automatic citation mapping tasks, the mapped citations constitute the "positive" data set while other citations in the Reference section that do not map to the given

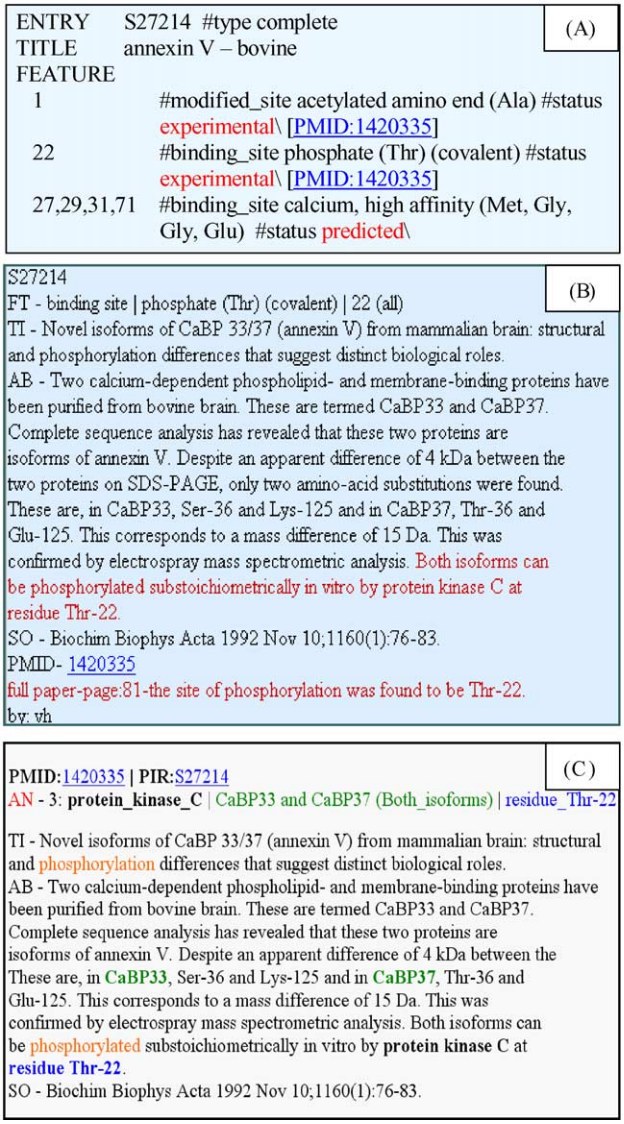

Fig. 3. Evidence attribution and computational extraction of experimental features in PIR-PSD. (A) Citation-attributed feature with PMID mapping; (B) annotation-tagged text based on manual curation; (C) annotation-tagged text based on computational extraction.

feature constitute the "negative" data set. For automatic annotation extraction tasks, the annotation-tagged texts can be used as training corpora for various types of protein sequence features, which are categorized and described based on a controlled vocabulary.

3.3. Training corpus for protein post-translational modifications (PTMs)

From the 9788 experimental feature lines, a total of 2037 lines correspond to five post-translational modification features—phosphorylation, acetylation, glycosylation, methylation, and hydroxylation. The status of citation mapping and evidence tagging of the five PTMs is shown in Table 1 (http://pir.georgetown.edu/cgi-bin/

Table 1
Citation mapping and evidence tagging of experimental PTM features in PIR-PSD

PTM types	# Feature lines	# Citation-mapped	# Evidence-tagged		
			AB	FL	NA
Acetylation	664	636 (95.7%)	79	107	401
Glycosylation	626	322 (51.4%)	121	74	136
Methylation	238	198 (83.1%)	38	36	107
Phosphorylation	303	255 (84.1%)	159	58	43
Hydroxylation	206	94 (45.6%)	41	32	35

PTM, post-translational modification; AB, abstract; FL, full-text article; NA, not tagged. (Data cited as of September 2004.)

ipkLitFt.pl?stat=2). The on-line table provides hypertext links for each PTM type to several underlying datasets, including the complete listing of all PSD entries with mapped citations (PMIDs) as well as the complete collection of evidence-tagged texts.

The PTM data sets can be exploited as NLP training and benchmarking data for identifying each of the five individual PTM types or, potentially, for the recognition and extraction of generic PTMs. The data are now being used for PTM annotation extraction by our collaborating computational groups. One example is the automatic extraction of protein phosphorylation information, including agent (kinases), substrate, and sites, from the abstract (Fig. 3C) using a rule-based system. Another example is the use of abstracts tagged for the five PTMs to train automatic classifiers to classify papers reporting PTMs based on support vector machine and Bayesian naïve statistical approaches. These studies allow "computer-assisted" retrospective literature survey and will facilitate literature-based feature annotation for protein databases.

Furthermore, such literature mining studies can benefit proteomic research because detecting protein PTMs (especially phosphorylation) is one of the major challenges in large-scale proteomic analyses. PTMs found in proteomes vary with cell and tissue types, and change in temporal manners. Literature mining techniques can assist the creation of a knowledgebase composed of experimental evidence for known PTMs that are mapped to protein entries. The knowledgebase can then serve as a useful reference for the identification and characterization of peptides from high-throughput proteomic data such as those from peptide mass fingerprints.

4. iProLINK resource for protein named entity recognition and ontology development

Protein named entity recognition (finding protein names from literature texts) is a prerequisite for bibliography mapping (identifying papers describing specified proteins). It is also fundamental for several other biological literature mining tasks, including the extraction of protein annotations (such as protein–protein interactions) from literature. Protein named entity recognition, however, is still an open problem and constitutes a bottleneck for computational mapping of bibliographic information to protein databases. The challenge primarily stems from the long-standing problem of protein nomenclature, where "profligate and undisciplined labeling is hampering communication" (Nature, 1997). A protein name is a label given to a protein object in the scientific literature and in biological databases. 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. Often the same protein is named differently in different databases or publications, and occasionally different proteins may share the same name. Protein name standardization requires community effort—only a small fraction of all proteins has standard nomenclature, most notably, the IUBMB Enzyme Nomenclature (http://www.chem.qmul.ac.uk/iubmb/enzyme).

There has been a small body of text mining work directly addressing the protein name problem (Fukuda et al., 1998; Yoshida et al., 2000; Zhou et al., 2004; Mika and Rost, 2004). The applications generally use three common approaches–dictionary-based, rule-based, and machine learning—and/or their combinations. The performance (precision and recall) of text mining techniques in biological name recognition remain relatively low (75–80%) compared to other domains. Multiple factors may be involved, including absence of shared training and test sets for rigorous measures of progress, lack of annotated training data specific to biological tasks, pervasive ambiguity of terms, frequent introduction of new terms, and a mismatch between evaluation tasks as defined for news report and for biological problems (Hirschman et al., 2002b). iProLINK consists of several data sources that can be used for protein named entity recognition.

4.1. Protein name dictionary and word token dictionaries

Our protein name dictionary is derived from the protein name field in the iProClass database (Wu et al., 2004b), which consists of protein names from UniProt (Swiss-Prot, TrEMBL, PIR-PSD) and RefSeq (Pruitt and Maglott, 2001). After the initial compilation, the dictionary undergoes several filtering processes to generate unique protein names (including synonyms and acronyms), and to remove nonsensical names and certain non-name annotations. For example, entry names such as "*Inter-alpha-trypsin inhibitor* (*GIK-14*) (*Fragment*)" were broken into *Inter-alpha-trypsin inhibitor*, *GIK-14* and *Fragment*. The name *Fragment* is later removed from the dictionary along with a list of other "bad" names such as *hypothetical protein*, *conserved hypothetical protein*, *unnamed protein product*, *predicted protein*, and *predicted protein of unknown function*. In addition, words such as *probable*, *putative*, and *similar to* before protein names are also removed so that a name like *putative aspartate aminotransferase A* is merged to *aspartate aminotransferase A* to reduce the redundancy. Derived from over 1.5 million iProClass

entries, the protein name dictionary currently has about 700,000 names, each of which is shown with its frequency count.

Most protein names are composed of combinations of two or more words (or tokens). Therefore, protein name rules can be derived from tokenized protein words and used during post-tagging processing to improve machine learning-based named entity recognition. We have compiled specialized single-word dictionaries by tokenization and classification of protein names from 30,000 well-curated iProClass protein entries (each containing at least five reference citations). The dictionaries consist of individual word tokens categorized into five classes:

- Biomedical terms (*bt*): these terms are used in a broad range of biological and medical sciences. They mainly describe structures of all forms of life at different levels (from gross morphology to molecular structure), as well as their respective functions and mechanisms in both normal (physiological) and diseased (pathological) states.
- Chemical terms (*ct*): these are words that describe organic or inorganic chemical materials, chemical groups or bonds, or chemical properties.
- Macromolecules (*mc*): these words refer to biopolymers such as proteins, peptides, DNA, RNA, polysaccharides, or glycoproteins.
- Common English (*ce*): common English words are used to describe various aspects or properties of proteins, such as *short*, *signal*, *interacting*, and *repair*. These also include spelled-out form of Greek letters, such as *alpha* and *beta*, as well as stop words like *of*, *at*, and *to*.
- Non-word tokens: they are combinations of letters, numbers, or symbols. They often are acronyms, synonyms, or abbreviations, such as *DNA* for *deoxyribonucleic acid*, *Ala* for *alanine*, and *GH* for *growth hormone*. The form of non-word tokens can be number only, single letter, multiple letters, or combinations of numbers, letters, and other symbols. Non-word tokens may stand for biochemical entities such as nucleic acids, nucleotides, and amino acids.

Protein name rules can be expressed based on the five token classes. Some examples include: (i) a protein name which is not an acronym or abbreviations should have at least one *mc* word, as in *natural killer cell-activating factor*, *parathyroid hormone receptor* 2, and *glutathione transferase* 4; (ii) *bt* and *ct* words alone cannot make protein names unless combined with an *mc* word, as in *transcription factor II*, *potassium channel*, and *nucleoside diphosphate phosphatase*; and (iii) numerals alone cannot be a protein name unless combined with other symbols, as in *p53*, *p38*, and *hsp70*.

4.2. Protein name tagging guidelines and name-tagged corpora

Other iProLINK data resources for named entity recognition are two sets of literature corpora that were manually tagged with protein names based on two versions of tagging guidelines. They were developed to test inter-annotator reliability and can be used as a gold standard for different machine-based protein name taggers or classifiers. Although several name-tagged gold standards already exist, the importance of inter-annotator reliability for machine-based taggers has been seldom addressed. In reality, due to the complex nature of protein naming, inter-annotator agreement varies. One can consider the inter-annotator performance as the upper-bound of the machine performance. To test inter-annotator reliability, each literature corpus of 300 abstracts was independently tagged by three individuals based on a common tagging guideline. Two protein name tagging guidelines (Versions 1.0 and 2.0) were developed and their effect on inter-annotator performance was compared (Mani et al., 2004). The major differences between the two tagging guidelines are summarized in Table 2.

Guideline 1.0 defines how to tag protein objects, not protein named entities. This leads to inconsistent tagging by different annotators when protein names refer to non-protein objects. Especially common inconsistency occurs when protein names are used in the context of gene-related objects (such as gene, promoter, and mRNA). For example, *photoreceptor G-protein alpha-subunit gene GNAT2* refers to a gene object, but *photoreceptor G-protein alpha-subunit* is a protein name that was tagged by one annotator, but not the other two. As a result, the human tagging only achieved a 0.716 *F*-measure among three annotators.

Guideline 2.0 defines tagging rules for protein named entities regardless of the context of the object. An exclusion list is given for generic terms such as *protein*, *subunit*, *activator*, and *carrier*. Thus, in the above example, *photoreceptor G-protein alpha-subunit* will be tagged regardless of what follows it. With the revised guideline, the second manually tagged data sets achieved a significantly higher inter-annotator *F*-measure of 0.868. Notably, an automatic name tagger based on our protein name dictionary alone tagged the same literature corpus with a performance of 0.41 *F*-measure based on the human-tagged corpus. Considering that many dictionary-based name tags overlap with the correct target entities as judged by humans, the machine tagging actually found 68% of the protein named entities in the 300 abstracts. Therefore, dictionary-based pre-tagging can facilitate the human tagging process by easing human reading and reducing human fatigue. In one example, the dictionary tagged multiple instances of *emerin* in an abstract, one or more of which was missed by each of the three annotators.

The confusion between protein objects and protein named entities is also observed in the entity extraction task of the BioCreAtIve contest. The goal of this task was to assess the ability of an automated system to identify genes (or proteins, where there is ambiguity) mentioned in text. It specifically required the identification of terms in biomedical articles that are gene or protein names. However, the distinction between a named entity and a gene/protein object was not explicit in the contest guideline. This ambiguity may have affected the performance of some tagging programs. Indeed,

Table 2
Comparisons of two versions of protein name tagging guidelines

	Tagging guideline 1.0	Tagging guideline 2.0
Tagging target	Protein object	Protein named entity
Tag types	<protein>, <acronym> and <array-protein>	<protein> and <long-form>
Use of dictionary	No dictionary	Pre-tagging with protein name dictionary
Prior knowledge	Major requirement	Minor requirement
Inter-annotator performance	*F*-measure: 0.716	*F*-measure: 0.868

several "false positives" tagged by one contested program (Liu et al., 2004) may have been legitimate "mentioning of genes/proteins." Some examples are *superoxide dismutase* in "... may be involved in copper homeostasis and modulation of copper/zinc *superoxide dismutase* (Cu/ZnSOD) activity in neurons" (PMID: 10550328), *TCR-delta* in "... and *TCR-delta* mRNA from lymph node" (PMID:10929051), and *Rad51* in "... arresting growth with S-phase DNA content, and generate nuclear *Rad51* foci, followed by cell death ..." (PMID: 11980714).

4.3. PIRSF family classification-based protein ontology

Biological ontologies are crucial for biological knowledge management, including mining literature data to extract relevant information and integrating information from multiple databases. A protein ontology—consisting of names and synonyms of protein classes as well as their relationships—can be used to assist with protein named entity recognition. Furthermore, an ontology based on protein family relationships, such as the PIRSF classification system, can be mapped to and complement the gene ontology (GO) (Ashburner et al., 2000).

The PIRSF (SuperFamily) classification system organizes proteins into a network structure from superfamilies to subfamilies to reflect evolutionary relationship of full-length proteins (Wu et al., 2004a). In the hierarchy, each protein is assigned to a single homeomorphic family where members share overall sequence similarity and common domain architecture. A homeomorphic family may have zero or more parent superfamilies and zero or more child subfamilies. The flexible number of parent–child levels from superfamily to subfamily reflects natural clusters of proteins with varying degrees of sequence conservation, rather than arbitrary similarity thresholds. We have developed a protein ontology based on PIRSF hierarchical family names. The ontology is in the GO flat file format (http://www.geneontology.org/GO.format.html) with a DAG (directed acyclic graph) structure, and can be browsed using application tools such as DAG-Edit (http://www.geneontology.org/GO.tools.html).

To evaluate how PIRSF can enrich GO concepts, we have mapped 5500 PIRSF homeomorphic families and subfamilies to the GO hierarchy. We found that about 67% of the curated PIRSF families and subfamilies map to GO leaf nodes, among which 2209 PIRSFs have shared GO leaf nodes. Our study indicates that a PIRSF-based protein ontology can complement GO in several ways. (1) It provides a mechanism to systematically examine the relationships between the three GO sub-ontologies (molecular function, biological process and cellular component) based on the shared annotations at different protein family levels. (2) The PIRSF associations to GO nodes can lead to interesting examinations as to whether certain GO subtrees might need expansion if GO concepts are too broad. (3) The comprehensive classification of related protein families in PIRSF can also suggest identification of missing GO nodes when entire groups of superfamilies or families cannot be mapped to existing GO terms.

5. Conclusion

We have developed iProLINK as a resource to facilitate text mining/NLP research in the areas of literature-based database curation, named entity recognition, and ontology development. The collection of data sources can be utilized by computational or biological researchers to explore literature information on proteins and their features or properties. Therefore, iProLINK serves as a knowledge link bridging protein databases and scientific literature.

Acknowledgements

The project is supported by grants ITR-0205470, DBI-0138188, and IIS-0430743 from the National Science Foundation and grant U01-HG02712 from the National Institutes of Health, USA.

References

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S., 2004. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 32, D115–D119.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Nat. Genet. 25, 25–29.

Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., Eppig, J.T., The members of the Mouse Genome Database Group, 2003. MGD: The mouse genome database. Nucleic Acids Res. 31, 193–195.

Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., Cherry, J.M., 2004. Saccharomyces genome database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. Nucleic Acids Res. 32, D311–D314.

Craven, M., Kumlien, J., 1999. Constructing biological knowledge bases by extracting information from text sources. Proc. Int. Conf. Intell. Syst. Mol. Bio. 7, 77–86.

Fukuda, K., Tamura, A., Tsunoda, T., Takagi, T., 1998. Toward information extraction: identifying protein names from biological papers. Pac. Symp. Biocomput, pp. 707–718.

Hirschman, L., Park, J.C., Tsujii, J., Wong, L., Wu, C.H., 2002a. Accomplishments and challenges in literature data mining for biology. Bioinformatics 18, 1553–1561.

Hirschman, L., Morgan, A.A., Yeh, A.S., 2002b. Rutabaga by any other name: extracting biological names. J. Biomed. Inform. 35, 247–259.

Liu, H., Wu, C.H., Friedman, C., 2004. BioTagger: a biological entity tagging system. BioCreative Workshop—A Critical Assessment of Text Mining Methods in Molecular Biology, Granada, Spain, March 28–31.

Mani, I., Hu, Z., Jang, S., Samuel, K., Krause, M., Phillips, J., Wu, C.H., 2004. Protein name tagging guidelines: lessons learned. In: Proceedings of BioLINK SIG, Intelligent Systems for Molecular Biology, Glasgow, pp. 1–5.

Mika, S., Rost, B., 2004. Protein names precisely peeled off free text. Bioinformatics 20 (Suppl 1), I241–I247.

Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M., Ward, J.M., 2003. Gene Indexing: characterization and analysis of NLM's GeneRIFs. Proc. AMIA Symp., 460–464.

Nature editorial, 1997. Obstacles of nomenclature. Nature 389, 1.

Pruitt, K.D., Maglott, D.R., 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29, 137–140.

Wu, C.H., Yeh, L.-S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J., Barker, W.C., 2003a. The Protein Information Resource. Nucleic Acids Res. 31, 345–347.

Wu, C.H., Huang, H., Yeh, L.S., Barker, W.C., 2003b. Protein family classification and functional annotation. Comput. Bio. Chem. 27, 37–47.

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.-S., Natale, D., Vinayaka, C.R., Hu, Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R.S., Suzek, B.E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J.L., Chung, S., Castro-Alvear, J., Dinkov, G., Barker, W.C., 2004a. PIRSF family classification system at the Protein Information Resource. Nucleic Acids Res. 32, D112–D114.

Wu, C.H., Huang, H., Nikolskaya, A., Hu, Z., Yeh, L.S., Barker, W.C., 2004b. The iProClass integrated database for protein functional analysis. Comput. Bio. Chem. 28, 87–96.

Yeh, A.S., Hirschman, L., Morgan, A.A., 2003. Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. Bioinformatics 19 (Suppl. 1), i331–i339.

Yoshida, M., Fukuda, K., Takagi, T., 2000. PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. Bioinformatics 16, 169–175.

Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C., 2004. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 20, 1178–1190.