Dynamically Generating a Protein Entity Dictionary Using Online Resources

Hongfang Liu
Department of Information Systems
University of Maryland, Baltimore County
Baltimore, MD 21250
hfliu@umbc.edu

Zhangzhi Hu **Cathy Wu**
Department of Biochemistry and Molecular Biology
Georgetown University Medical Center
3900 Reservoir Road, NW, Washington, DC 20057
{zh9,wuc}@georgetown.edu

Abstract: With the overwhelming amount of biological knowledge stored in free text, natural language processing (NLP) has received much attention recently to make the task of managing information recorded in free text more feasible. One requirement for most NLP systems is the ability to accurately recognize biological entity terms in free text and the ability to map these terms to corresponding records in databases. Such task is called biological named entity tagging. In this paper, we present a system that automatically constructs a protein entity dictionary, which contains gene or protein names associated with UniProt identifiers using online resources. The system can run periodically to always keep up-to-date with these online resources. Using online resources that were available on Dec. 25, 2004, we obtained 4,046,733 terms for 1,640,082 entities. The dictionary can be accessed from the following website: http://biocreative.ifsm.umbc.edu/biothesaurus/.
Contact: hfliu@umbc.edu

1 Introduction

With the use of computers in storing the explosive amount of biological information, natural language processing (NLP) approaches have been explored to make the task of managing information recorded in free text more feasible *[1, 2]*. One requirement for NLP is the ability to accurately recognize terms that represent biological entities in free text. Another requirement is the ability to associate these terms with corresponding biological entities (i.e., records in biological databases) in order to be used by other automated systems for literature mining. Such task is called biological entity tagging. Biological entity tagging is not a trivial task because of several characteristics associated with biological entity names, namely: synonymy (i.e., different terms refer to the same entity), ambiguity (i.e., one term is associated with different entities), and coverage (i.e., entity terms or entities are not present in databases or knowledge bases).

Methods for biological entity tagging can be categorized into two types: one is to use a dictionary and a mapping method *[3-5]*, and the other is to markup terms in the text according to contextual cues, specific verbs, or machine learning *[6-10]*. The performance of biological entity tagging systems using dictionaries depends on the coverage of the dictionary as well as mapping methods that can handle synonymous or ambiguous terms. Strictly speaking, tagging systems that do not use dictionaries are not biological entity tagging but biological term tagging, since tagged terms in text are not associated with specific biological entities stored in databases. It requires an additional step to map terms mentioned in the text to records in biological databases in order to be automatically integrated with other system or databases. Due to the dynamic nature associated with the molecular biology domain, it is critical to have a comprehensive biological entity dictionary that is always up-to-date.

In this paper, we present a system that constructs a large protein entity dictionary, BioThesaurus, using online resources. Terms in the dictionary are then curated based on high ambiguous terms to flag nonsensical terms (e.g., *Novel protein*) and are also curated based on the semantic categories acquired from the UMLS to flag descriptive terms that associate with other semantic types other than gene or proteins (e.g., terms that refer to species, cells or other small molecules). In the following, we first provide background and related work on dictionary construction using online resources. We then present our method on constructing the dictionary.

2 Resources

The system utilizes several large size biological databases including three NCBI databases (GenPept [11], RefSeq [12], and Entrez GENE [13]), PSD database from Protein Information Resources (PIR) [14], and

Proceedings of the ACL Interactive Poster and Demonstration Sessions,
pages 17–20, Ann Arbor, June 2005.

UniProt [15]. Additionally, several model organism databases or nomenclature databases were used. Correspondences among records from these databases are identified using the rich cross-reference information provided by the iProClass database of PIR [14]. The following provides a brief description of each of the database.

PIR Resources – There are three databases in PIR: the Protein Sequence Database (PSD), iProClass, and PIR-NREF. PSD database includes functionally annotated protein sequences. The iProClass database is a central point for exploration of protein information, which provides summary descriptions of protein family, function and structure for all protein sequences from PIR, Swiss-Prot, and TrEMBL (now UniProt). Additionally, it links to over 70 biological databases in the world. The PIR-NREF database is a comprehensive database for sequence searching and protein identification. It contains non-redundant protein sequences from PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB.

UniProt – UniProt provides a central repository of protein sequence and annotation created by joining Swiss-Prot, TrEMBL, and PSD. There are three knowledge components in UniProt: Swissprot, TrEMBL, and UniRef. Swissprot contains manually-annotated records with information extracted from literature and curator-evaluated computational analysis. TrEMBL consists of computationally analyzed records that await full manual annotation. The UniProt Non-redundant Reference (UniRef) databases combine closely related sequences into a single record where similar sequences are grouped together. Three UniRef tables UniRef100, UniRef90 and UniRef50) are available for download: UniRef100 combines identical sequences and sub-fragments into a single UniRef entry; and UniRef90 and UniRef50 are built by clustering UniRef100 sequences into clusters based on the CD-HIT algorithm [16] such that each cluster is composed of sequences that have at least 90% or 50% sequence similarity, respectively, to the representative sequence.

NCBI resources – three data sources from NCBI were used in this study: GenPept, RefSeq, and Entrez GENE. GenPept entries are those translated from the GenBanknucleotide sequence database. RefSeq is a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms. Entrez GENE provides a unified query environment for genes defined by sequence and/or in NCBI's Map Viewer. It records gene names, symbols, and many other attributes associated with genes and the products they encode.

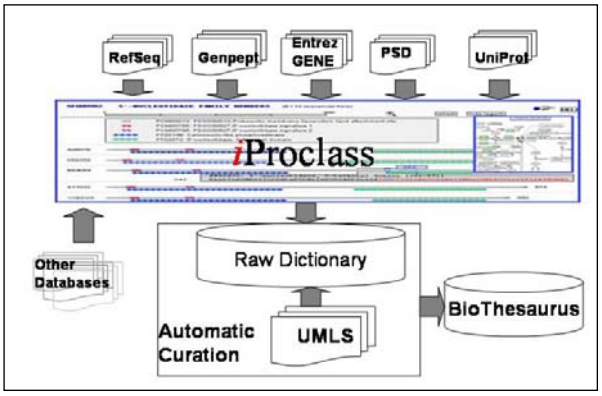


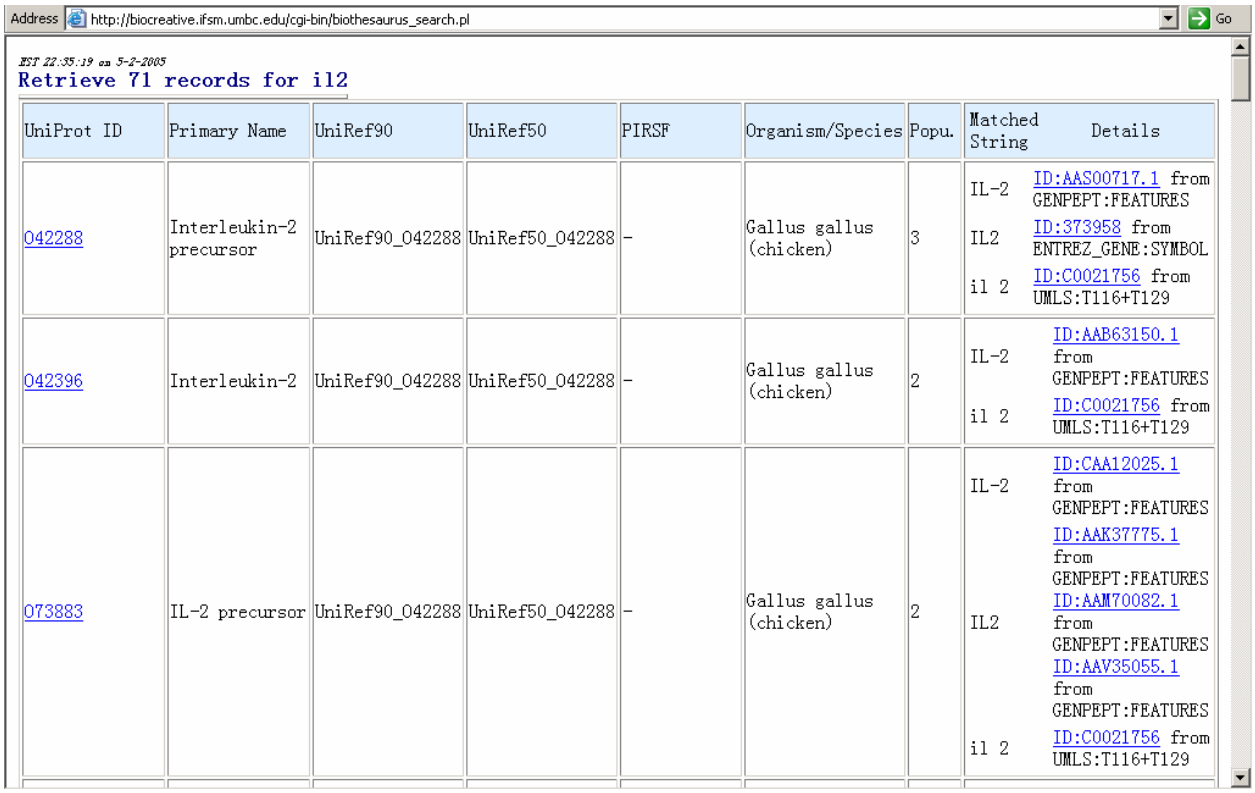
Figure 1: The overall architecture of the system

The UMLS – the Unified Medical Language System (UMLS) has been developed and maintained by National Library of Medicine (NLM) [17]. It contains three knowledge sources: the Metathesaurus (META), the SPECIALIST lexicon, and the Semantic Network. The META provides a uniform, integrated platform for over 60 biomedical vocabularies and classifications, and group different names for the same concept. The SPECIALIST lexicon contains syntactic information for many terms, component words, and English words, including verbs, which do not appear in the META. The Semantic Network contains information about the types or categories (e.g., "Disease or Syndrome", "Virus") to which all META concepts have been assigned.

Other molecular biology databases - We also included several model organism databases or nomenclature databases in the construction of the dictionary, i.e., mouse - Mouse Genome Database (MGD) [18], fly - FlyBase [19], yeast - Saccharomyces Genome Database (SGD) [20], rat – Rat Genome Database (RGD) [21], worm – WormBase [22], Human Nomenclature Database (HUGO) [23], Online Mendelian Inheritance in Man (OMIM) [24], and Enzyme Nomenclature Database (ECNUM) [25, 26].

3 System Description and Results

The system was developed using PERL and the PERL module Net::FTP. Figure 1 depicts the overall architecture. It automatically gathers fields that contain annotation information from PSD, RefSeq, Swiss-Prot, TrEMBL, GenBank, Entrez GENE, MGI, RGD, HUGO, ENCUM, FlyBase, and WormBase for each iProClass record from the distribution website



UniProt ID	Primary Name	UniRef90	UniRef50	PIRSF	Organism/Species	Popu.	Matched String	Details
042288	Interleukin-2 precursor	UniRef90_042288	UniRef50_042288	-	Gallus gallus (chicken)	3	IL-2 IL2 il 2	ID:AAS00717.1 from GENPEPT:FEATURES ID:373958 from ENTREZ_GENE:SYMBOL ID:C0021756 from UMLS:T116+T129
042396	Interleukin-2	UniRef90_042288	UniRef50_042288	-	Gallus gallus (chicken)	2	IL-2 il 2	ID:AAB63150.1 from GENPEPT:FEATURES ID:C0021756 from UMLS:T116+T129
073883	IL-2 precursor	UniRef90_042288	UniRef50_042288	-	Gallus gallus (chicken)	2	IL-2 IL2 il 2	ID:CAA12025.1 from GENPEPT:FEATURES ID:AAK37775.1 from GENPEPT:FEATURES ID:AAM70082.1 from GENPEPT:FEATURES ID:AAV35055.1 from GENPEPT:FEATURES ID:C0021756 from UMLS:T116+T129

Figure 2: Screenshot of retrieving il2 from BioThesaurus

of each resource. Annotations extracted from each resource were then processed to extract terms where each term is associated with one or more UniProt unique identifiers and comprised the raw dictionary for BioThesaurus. The raw dictionary was computationally curated using the UMLS to flag the UMLS semantic types and remove several high frequent nonsensical terms. There were a total of 1,677,162 iProclass records in the PIR release 59 (released on Dec 25 2004). From it, we obtained 4,046,733 terms for 1,640,082 entities. Note that about 27,000 records have no terms in the dictionary mostly because they are new sequences and have not been annotated and linked to other resources or terms associated with them are nonsensical. The dictionary can be searched through the following URL: http://biocreative.ifsm.umbc.edu/biothesaurus/Biothesaurus.html.

Figure 2 shows a screenshot when retrieving entities associated with term il2. It indicates that there are totally 71 entities in UniProt that il2 represents when ignoring textual variants. The first column of the table is UniProt ID. The primary name is shown in the second column, the family classifications available from iProClass are shown in the following several columns, the taxonomy information is shown in the next. The popularity of the term (i.e., the number of databases that contain the term or its variants) is shown next. And the last column shows the links to the records from which the system extracted the terms.

4 Discussion and Conclusion

We demonstrated here a system which generates a protein entity dictionary dynamically using online resources. The dictionary can be used by biological entity tagging systems to map entity terms mentioned in the text to specific records in UniProt.

Acknowledgements

The project was supported by IIS-0430743 from the National Science Foundation.

Reference

1. Hirschman L, Park JC, Tsujii J, Wong L, Wu CH: **Accomplishments and challenges in literature data mining for biology**. *Bioinformatics* 2002, **18**(12):1553-1561.

2. Shatkay H, Feldman R: **Mining the biomedical literature in the genomic era: an overview**. *J Comput Biol* 2003, **10**(6):821-855.
3. Krauthammer M, Rzhetsky A, Morozov P, Friedman C: **Using BLAST for identifying gene and protein names in journal articles**. *Gene* 2000, **259**(1-2):245-252.
4. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression**. *Nat Genet* 2001, **28**(1):21-28.
5. Hanisch D, Fluck J, Mevissen HT, Zimmer R: **Playing biology's name game: identifying protein names in scientific text**. *Pac Symp Biocomput* 2003:403-414.
6. Fukuda K, Tamura A, Tsunoda T, Takagi T: **Toward information extraction: identifying protein names from biological papers**. *Pac Symp Biocomput* 1998:707-718.
7. Sekimizu T, Park HS, Tsujii J: **Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts**. *Genome Inform Ser Workshop Genome Inform* 1998, **9**:62-71.
8. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K: **A biological named entity recognizer**. *Pac Symp Biocomput* 2003:427-438.
9. Tanabe L, Wilbur WJ: **Tagging gene and protein names in biomedical text**. *Bioinformatics* 2002, **18**(8):1124-1132.
10. Lee KJ, Hwang YS, Kim S, Rim HC: **Biomedical named entity recognition using two-phase model based on SVMs**. *J Biomed Inform* 2004, **37**(6):436-447.
11. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank: update**. *Nucleic Acids Res* 2004, **32 Database issue**:D23-26.
12. Pruitt KD, Katz KS, Sicotte H, Maglott DR: **Introducing RefSeq and LocusLink: curated human genome resources at the NCBI**. *Trends Genet* 2000, **16**(1):44-47.
13. NCBI: **Entrez Gene**. In., vol. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene; 2004.
14. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE *et al*: **The Protein Information Resource**. *Nucleic Acids Res* 2003, **31**(1):345-347.
15. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **UniProt: the Universal Protein knowledgebase**. *Nucleic Acids Res* 2004, **32 Database issue**:D115-119.
16. Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein databases**. *Bioinformatics* 2001, **17**(3):282-283.
17. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic Acids Res* 2004, **32 Database issue**:D267-270.
18. Bult CJ, Blake JA, Richardson JE, Kadin JA, Eppig JT, Baldarelli RM, Barsanti K, Baya M, Beal JS, Boddy WJ *et al*: **The Mouse Genome Database (MGD): integrating biology with the genome**. *Nucleic Acids Res* 2004, **32 Database issue**:D476-481.
19. Consortium F: **The FlyBase database of the Drosophila genome projects and community literature**. *Nucleic Acids Res* 2003, **31**(1):172-175.
20. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: **SGD: Saccharomyces Genome Database**. *Nucleic Acids Res* 1998, **26**(1):73-79.
21. Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen CF, Nigam R, Kwitek A *et al*: **Rat Genome Database (RGD): mapping disease onto the genome**. *Nucleic Acids Res* 2002, **30**(1):125-128.
22. Harris TW, Chen N, Cunningham F, Tello-Ruiz M, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Chan J *et al*: **WormBase: a multi-species resource for nematode biology and genomics**. *Nucleic Acids Res* 2004, **32 Database issue**:D411-417.
23. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H: **The HUGO Gene Nomenclature Committee (HGNC)**. *Hum Genet* 2001, **109**(6):678-680.
24. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2005, **33 Database Issue**:D514-517.
25. Gegenheimer P: **Enzyme nomenclature: functional or structural?** *Rna* 2000, **6**(12):1695-1697.
26. Tipton K, Boyce S: **History of the enzyme nomenclature system**. *Bioinformatics* 2000, **16**(1):34-40.