Mapping Gene/Protein Names in Free Text to Biomedical Databases

Hongfang Liu[1], Manabu Torii[1], Zhang-zhi Hu[2], and Cathy Wu[2]
[1]Department of Biostatistics, Bioinformatics & Biomathematics
[2]Protein Information Resource
Georgetown University Medical Center
{hl224, mt352, zh9, wuc}@georgetown.edu

Abstract

Observing that many biomedical databases have been developed and maintained independently, their records referring to the same entities may have different sets of synonyms. Integration of names pertaining to the same entity would provide a more comprehensive list of synonyms than each individual database. We have assembled BioThesaurus, a thesaurus of proteins and their corresponding genes compiled from multiple databases for all UniProtKB records. In this study, the coverage of BioThesaurus, and the contribution of each individual database were assessed for several organisms. The result indicates that the coverage of BioThesaurus is over 80% for most of the organisms with an average of 85.4%. When restricted to individual databases or resources, the percentages dropped ranging from 3 to 30%. The study demonstrated that each individual database or resource has some synonyms not covered by other databases or resources, and a list of names compiled from multiple databases would be desired for systems requiring high recall.

1. Introduction

As the pace of biomedical research accelerates, researchers become more and more dependent on computers to manage the explosive amount of biomedical information being published. Hundreds of biomedical databases have been generated to store the information [1]. Most of these databases were constructed and maintained independently, in which cross-references to other databases as well as to relevant literature are usually included to facilitate data integration [2]. The high quality of many databases is assured by database curators who extract and synthesize information stored in literature or other databases. Such manual process is very time-consuming. Natural language processing (NLP) techniques have been exploited to accelerate the curation process by retrieving literature and extracting information [3-5]. One prerequisite for NLP is to accurately recognize biological entity names in literature and map the identified names to corresponding records in biomedical databases [6]. Usually, a biomedical database provides a list of names either entered by curators or extracted from other databases. Those names could be used for mapping of a term in text to database records by NLP systems. Observing biomedical databases have been developed and maintained by different groups independently, names provided for records referring to a common entity may be different among them, and integrating names pertaining to a common entity together would provide a more comprehensive list of names for that entity than each individual database.

Several groups have developed resources by integrating names from multiple resources. For example, GENA automatically gathers official gene symbols, official full names, and synonyms from several databases, such as Swiss-Prot, FlyBase, and MGD [7]. ProMiner extracted gene symbols, alias names, and full names from HUGO, Swiss-Prot and TrEMBL [8]. Both GENA and ProMiner were developed primarily for biological named entity tagging systems and names were not directly linked to records in biological databases. GPSDB is another database that collects names from more than a dozen of biomedical databases for the purpose of query expansion when retrieving papers from MEDLINE [9]. Utilizing rich cross-references provided by iProClass and UniProtKB, we generated BioThesaurus, a thesaurus of gene and protein names for all records in UniProtKB, which has been incorporated into the PIR web search service to provide mapping between names and UniProtKB records [10].

0-7695-3019-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDMW.2007.94

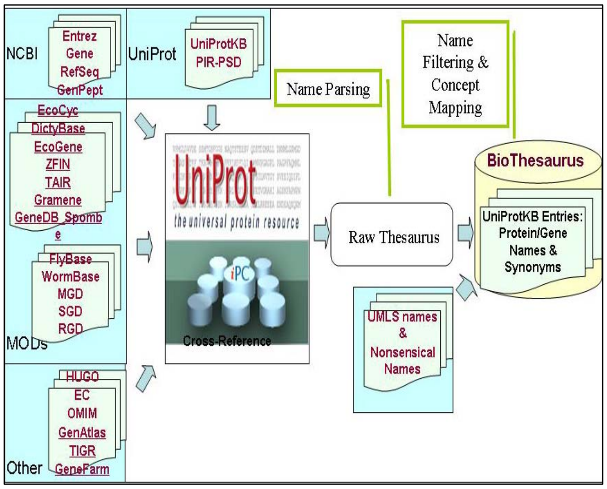

Figure 1. An overview of BioThesaurus construction (34 databases were used to compile BioThesaurus).

Table 1. Summary of model organisms under study.

Org.	Tid	Organism-Specific Database
E.Coli	562	EcoGene: http://www.ecogene.org EcoCyc: http://www.ecocyc.org
Th.Cr	3702	TAIR: http://www.arabidopsis.org/
Yeast	4932	SGD: http://www.yeastgenome.org/
Worm	6239	WormBase: http://www.wormbase.org
Fly	7227	FlyBasehttp://flybase.bio.indiana.edu/
Human	9606	HGNC: http://www.gene.ucl.ac.uk/nomenclature/
Mouse	10090	MGI: http://www.informatics.jax.org/
Rat	10116	RGD: http://rgd.mcw.edu/

In this paper, we used BioThesaurus to compare the coverage of a list of names provided by each individual database to the coverage of a list of names compiled from multiple databases. The study chose several organisms including human, mouse, yeast, etc. It utilized a data set automatically generated based on literature cross-references provided by UniProtKB. The data set consists of pairs (R, C), where R is a UniProtKB record associated with the list of organisms and C is a MEDLINE citation cross-referenced by R. We also investigated the contribution of databases when generating a list of names used for mapping names to UniProtKB records.

In the following, we first present background information on BioThesaurus. The method used in the study is presented next. We then provide a discussion and conclude our work.

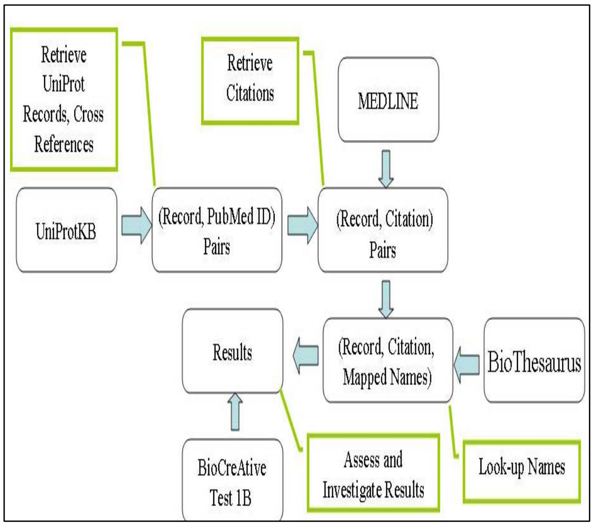

Figure 2. An overall architecture of the study.

2. Construction of BioThesurus

Figure 1 shows an overview of the construction of BioThesaurus. The detail description of BioThesaurus can be found in our previous paper [11]. BioThesaurus was intended to provide comprehensive protein and gene names for all protein records in UniProtKB knowledgebase (UniProtKB). Names in BioThesaurus were extracted from multiple biomedical databases based on the cross-references provided by two protein information databases: iProClass and UniProtKB. A total of thirty four underlying databases were used to construct BioThesaurus used in this paper (version 4, June 2007) and also summarized in the following: i) protein databases maintained by UniProtKB, including UniProtKB and PIR-PSD, ii) gene and protein resources at NCBI, including Entrez Gene, RefSeq, and GenPept (GenBank translation), iii) organism-specific-databases such as MGD, SGD, RGD, FlyBase, ZFIN, EcoGene, EcoCyc, TAIR, and WormBase, and iv) a few other databases, such as HUGO, GenAtlas, the EC enzyme nomenclature, BioCyc, and the OMIM database of human genes and genetic disorders. BioThesaurus is available at the PIR iProLINK website[1].

3. Methods

In this study, we chose human and several model organisms based on to the criteria that each organism should have a corresponding model organism database.

[1] http://pir.georgetown.edu/iprolink

Table 1 shows organisms that were included in the study where the common name, NCBI TAXON identifier (Tid), and corresponding organism-specific database(s) are shown. Figure 2 shows the overall architecture of the study. For each organism, we identified its protein records in UniProtKB and retrieved their corresponding MEDLINE cross-references. The corresponding MEDLINE citations were then retrieved from PubMed. We looked up names associated with the records in these citations and assessed the results. All online resources used here were accessed on June 2007.

3.1. Retrieve UniProtKB Records and Cross-references

We retrieved UniProtKB records associated with each organism. For example, based on the cross-reference information provided by UniProtKB, we associated six PubMed identifiers (PMIDs) (i.e., 10025402, 10196122, 10748113, 3317403, 3344209, 7532276) with UniProtKB record P63012, RAB3A_RAT.

3.2. Retrieve MEDLINE citations

Based on PubMed identifiers, we used BATCH ENTREZ[2] to retrieve MEDLINE citations. For example, citation identified by PMID 10025402 as shown in Figure 3 was retrieved and associated with UniProtKB record P63012 (i.e., RAB3A_RAT). Note that the relationship between records and citations is many-to-many. One record can be associated with multiple citations and one citation may be associated with multiple records. As indicated before, there were six citations associated with UniProtKB record P63012. Among them, five were associated with multiple records. For example, the citation shown in Figure 3 was also associated with record P47709 (i.e., RP3A_RAT). We will denote the association between record R and citation C as a pair (R, C).

3.3. Look up names

For each pair (R, C), we retrieved names for R in BioThesaurus and looked them up in the title and abstract of C. We employed two ways to look up names in MEDLINE citations: exact and flexible. In the first approach, exact string matching is used, while in the latter approach, case difference, lexical difference and punctuation difference are ignored during the look-up procedure. For example, *rabphilin-3A,* a name for protein record P47709, is present in both the title and abstract of the citation in Figure 3, and they can be identified using exact string matching. However, *Rab3a*, a name for P63012, fails to be mapped using exact string matching, since *Rab3a* does not occur in the citation but its case variant, *Rab3A* is present. The occurrence of names for P63012 could be identified when using flexible matching.

ID – 10025402

TI - Structural basis of Rab effector specificity: crystal structure of the small G protein Rab3A complexed with the effector domain of rabphilin-3A.

AB - The small G protein Rab3A plays an important role in the regulation of neurotransmitter release. The crystal structure of activated Rab3A/GTP/Mg2+ bound to the effector domain of rabphilin-3A was solved to 2.6 A resolution. Rabphilin-3A contacts Rab3A in two distinct areas. The first interface involves the Rab3A switch I and switch II regions, which are sensitive to the nucleotide-binding state of Rab3A. The second interface consists of a deep pocket in Rab3A that interacts with a SGAWFF structural element of rabphilin-3A. Sequence and structure analysis, and biochemical data suggest that this pocket, or Rab complementarity-determining region (RabCDR), establishes a specific interaction between each Rab protein and its effectors. RabCDRs could be major determinants of effector specificity during vesicle trafficking and fusion.

Figure 3. A sample citation 10025402 retrieved for protein records: P63012 and P47709.

Table 2. Statistics of records (R), MEDLINE citations (C), and pairs (R, C) extracted for each organism.

	# (R.,C)	# C	# R
E.Coli	34,892	7,239	13,658
Th. cr	32,060	3,999	20,068
Yeast	32,906	7,975	7,105
Worm	25,874	1,508	23,041
Fly	92,201	4,338	21,946
Human	133,100	45,115	58,220
Mouse	225,316	17,625	54,279
Rat	19,717	10,030	12,076
Total	596,066	92,029	210,393

[2] http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=PubMed

Table 3. Assessment results with respect to five ranges for eight organisms for two different mapping methods.

	E.Coli	Th.Cress	Yeast	Worm	Fly	Human	Mouse	Rat	Total
I **[1-5]**	10,332 (74.6) *35.0*	5,707 (82.1) *70.5*	11,166 (84.9) *60.0*	2,175 (83.3) *66.3*	5,496 (90.5) *76.2*	55,123 (87.6) *74.4*	21,981 (86.0) *55.4*	12,079 (83.5) 57.7	124,059 (85.4) *64.6*
II **[6-10]**	1,862 (48.1) *18.2*	1,159 (53.2) *37.4*	1,670 (50.7) *29.6*	198 (70.7) *44.9*	623 (67.9) *45.4*	4,668 (48.8) *41.5*	1,597 (43.1) *23.4*	610 (46.4) *23.8*	12,417 (49.7) *33.0*
III **[11-20]**	2,139 (25.4) *11.2*	912 (37.4) *21.2*	1,778 (37.5) *27.3*	186 (35.5) *18.3*	375 (23.7) *20.0*	4,427 (46.2) 42.0	1,092 (29.1) 11.9	340 (16.8) *3.2*	11,249 (36.7) *26.9*
IV **[21-40]**	1,328 (22.7) *6.6*	1,478 (40.7) *21.4*	1,299 (21.1) *18.5*	106 (36.8) *0*	382 (40.8) *27.7*	4,881 (34.5) *30.3*	628 (24.7) *13.4*	57 (0) *0*	10,159 (31.6) *22.7*
V **[>40]**	19,229 (2.9) *0.9*	22,804 (1.3) *0.9*	16,993 (4.0) *3.0*	23,209 (<0.1) *<0.1*	85,325 (1.3) *1.7*	63,994 (5.9) 5.0	200,011 (0.3) *0.2*	6,630 (2.2) *1*	438,195 (1.6) *1.4*

3.4. Assess the mapping results

We counted the number of pairs (R, C), where a name in BioThesaurus associated with R was found in the title and abstract sections of C with respect to two ways of matching (i.e., exact and flexible). Noticing some citations provided high throughput genome wide studies and have been cross-referenced by hundreds of records, we divided the pairs into five ranges: I (1-5), II (6-10), III (11-20), IV (21-40), and V (40 or more) where the range inside parentheses refers to the number of Rs that C refers to. For example, if a pair (R, C) is in range I, which indicates the number of UniProtKB records cross-referenced by C is less than six. For pairs in range I, we compared the coverage of (i) names obtained using NCBI resources only (denoted as NCBI), (ii) names provided by organism-specific databases only (denoted as MOD), and (iii) names in UniProtKB only (denoted as UniProtKB). Since one name can occur in multiple resources, we also assessed the absolute gain of each individual resource.

3.5. Investigate the coverage

For pairs (R,C) where C was associated with less than six records, we investigated the possibility that some synonyms for R occurring in the title and abstract of C were not captured by BioThesaurus, that is, there is a name for protein record R occurring in the title or abstract sections of C, but the name fails to be included by curators. We used expert-annotated corpora available at University of Colorado[3] to investigate missing synonyms[11]. We used the following corpora, in which protein names are annotated in the title and abstract sections: Bio1[4] , PIR[5] , GENIA[6] , and Yapex[7].

We combined citations from these annotated corpora and formed a list of expert-annotated citations. For pairs (R, C) in range I where none of the names for R could be mapped in C, if C is also in the obtained list, we reviewed names obtained from annotated citation(s) corresponding to C, identified potential causes, and checked potential synonyms.

4. Results

There are totally 210,393 UniProtKB records containing cross-references to MEDLINE, with a total of 92,029 unique referenced citations. Table 2 shows the number of records and citations retrieved for each organism. For example, we retrieved 45,115 citations for 58,220 human protein records in UniProtKB, with a total of 133,100 association pairs (R, C).

Tables 3 and 4 show the assessment results. In Table 3, the first number in each cell is the total number of pairs in that range. The second number (in parentheses) is the percentage of pairs (R, C) that could be mapped when flexible matching is used. The third number (shown in italic font) is the percentage when performing exact matching. For example, there are totally 124,059 pairs in range I. Names associated with R for 85.4% of them were found in C, when considering exact matching, the percentage drops to 64.6%. The results about the coverage of names for pairs in range I when limited to NCBI resources, organism-specific databases, and UniProtKB are

[3] http://compbio.uchsc.edu/corpora/obtaining.shtml

[4] http://research.nii.ac.jp/~collier/resources/bio1.1.xml

[5] http://pir.georgetown.edu/iprolink

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

[7] http://www.sics.se/humle/projects/prothalt/

Table 4. Assessment results with respect to names from individual databases.

	E.Coli	Th.Cress	Yeast	Worm	Fly	Human	Mouse	Rat
BioThesaurus	7,709 (74.6)	4,685 (82.1)	9,484 (84.9)	1,811 (83.3)	4,975 (90.5)	48,278 (87.6)	18,894 (86.0)	10,091 (83.5)
MOD	5,158/364 (49.9)	3,036/88 (53.2)	8,750/160 (78.4)	1,257/18 (57.8)	4,758/243 (86.6)	39,102/828 (70.9)	15,264/163 (69.4)	3,114/237 (25.8)
NCBI	6,758/344 (65.4)	4,461/530 (78.2)	8,901/243 (79.7)	1,730/146 (79.5)	4,577/52 (83.3)	45,789/3,763 (83.1)	18,040/1,052 (82.1)	8,943/1,639 (74.0)
UniProtKB	6,901/560 (66.8)	3,857/185 (67.6)	9,039/248 (81.0)	1,658/79 (76.2)	3,991/48 (72.6)	40,201/1,921 (72.9)	16,067/814 (73.1)	8,054/1,134 (66.7)

shown in Table 4. The number in parentheses shows the percentage and the number after slash shows the absolute contribution when including them in BioThesaurus. For example, among 55,123 pairs in human, there are 48,278 (87.6%) pairs (R, C) with names in BioThesaurus for R found in the title and abstract of C. The numbers reduced to 39,102 (70.9%), 45,789 (83.1%), and 40,201 (72.9%) when only considering names from HGNC, NCBI, and UniProtKB, respectively. The absolute contribution for HGNC is 828 (1.5% of 55,123). That is, if we exclude names in HGNC from mapping, the number of pairs (R, C) with names in BioThesaurus for R found in C would be 47,450 (i.e., reduced by 828). Figure 4 translates the numbers in tables 3 and 4 into charts. The first chart compares the coverage between two different mapping methods: flexible and exact. We can see that flexible mapping increases the coverage significantly (greater than 10% gain). The second chart shows the relative coverage of individual databases (i.e., MOD, NCBI, or UniProtKB) comparing to BioThesaurus. And the third chart shows the relative coverage of BioThesaurus when ignoring names from individual databases. We can see that the relative coverage of each individual database ranges from 30 to 95 percents. The absolute contribution of each individual database or resource is a range of 3 to 20%.

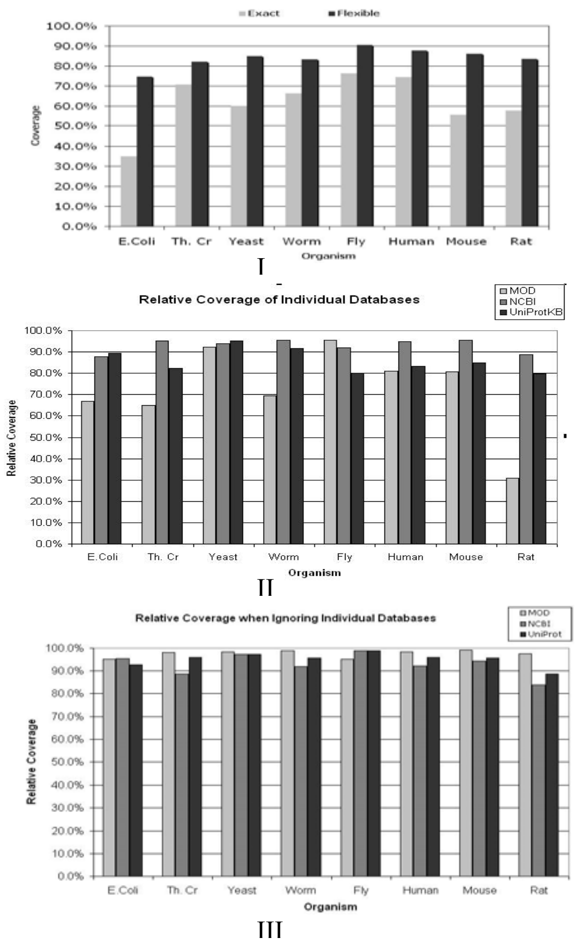

Figure 4. Assessment charts.

We extracted totally 2,803 annotated instances from four corpora, with 2,775 unique citations. There were totally 236 overlapped citations between the list of expert-annotated citations and citations in range I with a total of 344 pairs. Among them, 316 (91.9%) pairs were identified using BioThesaurus. For 28 pairs failed to be identified, eight of the pairs (R, C) failed to be mapped because of missing synonyms in BioThesaurus. Some of these synonyms could be automatically generated. For example, in pair (Q81U60, 12417715), the synonym *hDcp2* occurring in text for Q81U60 could be automatically generated by associating the initial of the corresponding organism common name (*h* for human) at the beginning of the symbol (*Dcp2*), 11 pairs were identified as cases where some general names (e.g., the protein family names) mentioned instead of specific names, and the remaining could not be mapped because names for record R were mentioned only in full text.

5. Discussion and Conclusion

We have presented a study about the coverage of names in BioThesaurus, which were compiled from various databases and names in free text using a data set that was automatically generated using cross-references provided in UniProtKB. We demonstrated in this paper that names in biomedical databases when combined could be mapped to names in the cross-referenced citations in UniProtKB with percentages ranged from 75 to 90% for several model organisms. Most of the names in BioThesaurus were present in multiple databases but there are some names unique to each database. From Table 4 and Figure 4, we can see that BioThesaurus always has the highest percentage of pairs (R, C) with names being matched to C, while the percentages assessed using names from individual databases could drop dramatically. We concluded that utilizing multiple resources to assemble synonyms for biological entities can have a very good coverage for names mentioned in text when using flexible matching.

We encountered 33 errors when retrieving citations using BATCH ENTREZ. A total of 11 PubMed identifiers were incorrect. For example, the citation (with PubMed identifier 11530255, titled *Cloning and characterization of a new soluble murine J-domain protein that stimulates BiP, Hsc70 and DnaK ATPase activity with different efficiencies*, in 2001 Gene, 273, 267-274) referred by UniProtKB record Q91ZF0 could not be retrieved. A close investigation shows that the correct PubMed identifier associated with the paper is 11595173. Note that some citations occur in several organisms and the number, 92,029 shown in the last row and the third column in Table 2 is the total number of unique citations. The organism with the most unique citations is human (a total of 45,115), and mouse has the most association pairs (a total of 225,316). There were 16 citations where each of them corresponds to more than ten thousands of UniProtKB records. For example, the citation (with PMID 16141072) referring to a paper by Carninci P et al., *The transcriptional landscape of the mammalian genome*, was associated with 30,416 mouse protein records.

From Table 3, we can see that if a citation C is cross-referred by many records, there is little chance names for those records could be found in the title and abstract of C. For example, the percentages for range I are around 80s while the percentages for range V are less than 2. For citations associated with a lot of records, by intuition, it is unlikely that names for specific records could be found in the title and/or citation of the abstract. The percentages obtained differ for each organism but not significantly. From Table 3, we also found that allowing flexible matching increases the chance of finding names in free text by over 10% in most of the cases. However, flexible matching imposes higher ambiguity to general English words (i.e., potential false positives during mapping, e.g., *CAT* vs. *cat*). In the future, we plan to use corpus-based word sense disambiguation to resolve the ambiguity between biological entity names and general English words.

The investigation of pairs with names failed to be found indicates that there are some synonyms in the text failed to be captured in BioThesaurus. Another future research direction would be to investigate ways to capture those names in free text automatically.

Acknowledgment: This work is supported by IIS-0639062 from the National Science Foundation.

REFERENCE

[1]. F. Bry and P. Kröger, "A Computational Biology Database Digest: Data, Data Analysis, and Data Management," Distributed and Parallel Databases, vol. 13, pp. 7-42, 2003.

[2]. W. Sujansky, "Heterogeneous database integration in biomedicine," J Biomed Inform, vol. 34, pp. 285-98, 2001.

[3]. L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, "Accomplishments and challenges in literature data mining for biology," Bioinformatics, vol. 18, pp. 1553-61, 2002.

[4]. H. Shatkay and R. Feldman, "Mining the biomedical literature in the genomic era: an overview," J Comput Biol, vol. 10, pp. 821-55, 2003.

[5]. W. Hersh, "Evaluation of biomedical text-mining systems: lessons learned from information retrieval," Brief Bioinform, vol. 6, pp. 344-56, 2005.

[6]. L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology," BMC Bioinformatics, vol. 6 Suppl 1, pp. S1, 2005.

[7]. A. Koike, Y. Kobatashi, and T. Takagi, "Kinase pathway database: An integrated protein-kinase and NLP-based protein-interaction resource," Genome Research, vol. 13, pp. 1231-1243, 2003.

[8]. D. Hanisch, J. Fluck, H. T. Mevissen, and R. Zimmer, "Playing biology's name game: identifying protein names in scientific text," Pac Symp Biocomput, pp. 403-14, 2003.

[9]. V. Pillet, M. Zehnder, A. K. Seewald, A. L. Veuthey, and J. Petrak, "GPSDB: a new database for synonyms expansion of gene and protein names," Bioinformatics, vol. 21, pp. 1743-4, 2005.

[10]. C. H. Wu, H. Huang, A. Nikolskaya, Z. Hu, and W. C. Barker, "The iProClass integrated database for protein functional analysis," Comput Biol Chem, vol. 28, pp. 87-96, 2004.

[11]. H. Liu, Z. Z. Hu, J. Zhang, and C. Wu, "BioThesaurus: a web-based thesaurus of protein and gene names," Bioinformatics, vol. 22, pp. 103-5, 2006.