Gene/protein/family name recognition in biomedical literature

Asako Koike [1,2]

[1]Central Research Laboratory, Hitatchi, Ltd.
1-280 Higashi-koigakubo Kokubunji, Tokyo, 185-8601
`akoike@hgc.jp`

Toshihisa Takagi[2]

[2]University of Tokyo. Dept. of Comp. Biol. Graduate School of Frontier Science Kiban-3A1(CB01) 1-5-1 Kashiwanoha Ka-shiwa-shi Chiba 277-8561, Japan
`tt@k.u-tokyo.ac.jp`

Abstract

Rapid advances in the biomedical field have resulted in the accumulation of numerous experimental results, mainly in text form. To extract knowledge from biomedical papers, or use the information they contain to interpret experimental results, requires improved techniques for retrieving information from the biomedical literature. In many cases, since the information is required in gene units, recognition of the named entity is the first step in gathering and using knowledge encoded in these papers. Dictionary-based searching is useful for retrieving biological information in gene units. However, since many genes in the biomedical literature are written using ambiguous names, such as family names, we need a way of constructing dictionaries. In our laboratory, we have developed a gene name dictionary:GENA and a family name dictionary. The latter contains ambiguous hierarchical gene names to compensate GENA. In addition, to address the problem of trivial gene name variations and polysemy, heuristics were used to search gene/protein/family names in MEDLINE abstracts. Using these algorithms to match dictionary and gene/protein/family names, about 95, 91, and 89% of protein/gene/family names in abstracts on *Saccharomyces cerevisiae, Drosophila melanogaster,* and *Homo sapiens* were detected with a precision of 96, 92, and 94%, in respective organisms. The effect of our gene/protein/family recognition method on protein-interaction and protein-function ex-traction using these dictionaries is also discussed.

1 Introduction

With the increasing number of biomedical papers, and their electronic publication in NCBI-PUBMED, there is a growing focus on information retrieval from texts. In particular, the recent development of procedures for large-scale experiments, such as yeast-two hybrid screening, mass spectrometry, and DNA/protein microarrays, has brought about many changes in the knowledge required by biologists and chemists. Because they produce large amounts of data on genes at one time, biologists require extensive knowledge of numerous genes to analyze the data obtained and these are beyond the capability of manual acquisition from the vast biomedical literature. Since, in many cases, the main objective of text processing is extraction of protein-protein/gene interaction or gene function, the first problem to solve is gene/protein/compound name recognition. To date, various methods of protein/gene name taggers have been proposed, mainly relating to *Homo sapiens*. These methods can be roughly divided into rule-based approaches (Fukuda et al. 1998), statistical approaches, including machine learning (Collier et al. 2000, Nobata et al. 1999), dictionary/knowledge-based approaches Humphreys et al. 2000, Jenssen et al. 2001, Koike et al. 2003), or a combination of these approaches (Tanabe and Wilbur, 2002). Since merely recognizing gene/protein names is insufficient to keep the extracted information in gene order, dictionary-based name recognition appears useful for assigning the locus of the extracted gene/protein name. Naming conventions are quite different for different organisms. Therefore, an appropriate approach is required for each organism.

There are three main problems in dictionary-based searching: (1) the existence of multi-sense words; (2) variations in gene names; and (3) the existence of ambiguous names. The first problem is mainly seen in symbol (abbreviated) types. For example, HAC1 is a synonym for both "tripartite motif-containing 3" and "hyperpolarization activated cyclic nucleotide-gated potassium channel 2" in *H. sapiens*. Further, some gene names, especially in *Drosophila melanogaster,* have the same spelling with verb(lack, ...), adjective(white, yellow...), common nouns (spot, twin, ...), and prepositions (of, ...). The second problem is trivial variations in gene names (orthographical, morphological, syntactic, lexicosemantic, insertion/deletion, permutation, or pragmatic). For example, "mitogen-activated protein kinase 1" and "protein kinase mitogen-activated, 1", "NIK serine/threonine protein kinase", and "NIK protein kinase" indicate the same gene. The third problem is caused by ambiguous expression of the gene name in the text. The problems of multi-sense words and the ambiguity are well summarized by Tuason *et al*. (2004)

In many cases, the family name is used instead of the gene name. A unique gene locus may not have been specified, especially for genes with multiple paralogs, or to avoid repeating the same expression, the family name may frequently be used. For example, in 1996, the "14-3-3" family name was counted 107 times in abstracts using mesh terms for *human*, while "14-3-3 alpha, beta, delta, gamma" gene name expressions did not appear at all. Thus, a family name dictionary is also required along with a gene name dictionary to specify the gene locus or loci. In this study, the above-mentioned problems were, as far as possible, solved simply using heuristics.

2   Construction of the gene name dictionary

The gene name dictionary, GENA, was constructed using the major databases, GenAtlas (http://www.dsi.univ-paris5.fr/genatlas/), HUGO (http://www.gene.ucl.ac.uk/hugo/), LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), GDB (http://gdb.weizmann.ac.il/index.shtml), SGD (http://www.yeastgenome.org/), MIPS (http://mips.gsf.de/genre/proj/yeast/index.jsp), Wormbase (http://www.wormbase.org/), OMIM (http://www.ncbi.nlm.nih.gov/omim/), MGI (http://www.informatics.jax.org/), RGD (http://rgd.mcw.edu/), FlyBase (http://flybase.bio.indiana.edu/), *S. pombe* geneDB (http://www.sanger.ac.uk/Projects/S_pombe/), SWISS-PROT, TrEMBL (http://us.expasy.org/sprot/), and PIR (http://pir.georgetown.edu/) for *Schizosaccharomyces*

pombe, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Mus musculus, Rattus norvegicus, and *Homo sapiens*, respectively. A merge of each database entry was done using the 'official symbol' or ORF name and link data provided by each entry and the protein-sequence data entry. The priority of the database was given in advance. For example, in *H. sapiens*, HUGO, Locuslink, GDB, and GenAtlas were registered in this order, using the merged entry for the same 'official symbol'. LocusLink's 'preferred symbol', which is not yet administered by HUGO, was also used. Merging the entries in SWISS-PROT, TrEMBL, and these registered data was done using the link data for 'Genew' provided by SWISS-PROT and TrEMBL. The rest of the entries were merged using the protein-IDs for LocusLink, SWISS-PROT, and TrEMBL. For example, LocusLink provides unique representative mRNA and protein sequences, and related sequences belonging to the same gene. If the protein-sequence entry for SWISS-PROT and TrEMBL matched with any of these sequence entries for LocusLink, the entries were merged. Linking these registered data with the PIR entries was also done using protein-ID entries. In principle, for all organisms, protein sequences without 'official or preferred symbols' were not registered. The entries consisted of 'official symbols' and 'official full names', which were provided by representative institutions, such as HUGO, for each organism, and 'synonyms' and 'gene products'. *S. cerevisiae* and *C. elegans* do not have 'official full names'. The distinction between these elements of each 'name' simply depends on the 'item headings' for each database. Although gene names and their product names are registered separately for one locus, and whether the entry's product is protein or RNA is also registered in GENA, we do not distinguish between them here. Hereafter, we do not distinguish 'gene product' from the gene name 'synonym'. Unfortunately, databases contain numerous mistakes or inappropriate gene/protein names. The reliability of each synonym was judged according to the database source. To meet our information extraction purposes, only gene names over a certain reliability can be used. Meaningless names (ex. hypothetical protein), higher concept names (ex. membrane protein) and apparently wrong names (ex. OK) were removed from the data semiautomatically using word-net vocabularies and term frequencies of all abstracts of one year. In an evaluation of this study, synonym names entered only in TrEMBL or PIR, except for names manually checked in our laboratory, were removed due to their low reliability.

In addition to these data, we added synonym names using the following methods. (1) Abbreviations of synonyms were added using an abbreviation extraction algorithm (Schwartz and Hears, 2003). (2) Plausible gene names were extracted from the subject and object noun of some verbs, which restricted such subjects and

objects as 'phosphorylate' and 'methylate' (both subjects must be protein/gene/family names). These are by-products of protein-interaction extraction in our project. The corresponding 'official symbol' was searched using a partial match of registered names, and finally was checked manually.

Compound names were gathered from the index of the biochemical dictionary, KEGG (http://www.genome.ad.jp/kegg/kegg2.html), mesh terms, and UMLS (http://www.nlm.nih.gov/research/umls/) and were registered in GENA. Some high-concept terms were removed manually. Compound name searches were not evaluated in this study. Currently (January, 2004), it contains about 920,000 registered gene/protein names and 210,000 compound names.

GENA was managed using Postgres, which provides command line searching and Web searching (http://www.gena.ontology.ims.u-tokyo.ac.jp). Searches can be done considering the word order replacement of long gene names using indexing all words consisting names.

3   Construction of family name dictionary

The construction of the family name dictionary was done using SWISS-PROT family names, PIR family names, INTERPRO family names (http://www.ebi.ac.uk/interpro/), gene/protein names in GENA, and clustering sequence similarities. These have hierarchical named entities. For example, "MAPK1" is a member of the "MAPK family" and the "MAPK family" is a member of the family of the "Ser/Thr protein kinase family"; in turn, this family is a member of "protein kinase", and "protein kinase" is a type of "kinase". Although "family" is usually used to indicate "similar sequence groups that probably have the same origin", sometimes it is also used to mean "sequence groups that have almost the same function". In this paper, we use "family" as "ambiguous gene/protein names that indicate similar sequences or biological functions". Plausible family names based on gene names are the common parts of multiple gene names, such as "MAPK" of "MAPK[number]", "14-3-3" of "14-3-3 [Greek alphabet[alpha-delta/alphabet[a-d]]", "protein kinase" of "Tyr protein kinase" and "Ser/Thr protein kinase", and "kinase" of "Inositol kinase" and "protein kinase". The backbone of the family hierarchy was constructed based on the INTERPRO family hierarchy. As far as possible, the remaining hierarchy was manually constructed considering sequence similarities, using Markov clustering (Enright et al. 2002) based on all-versus-all blast. The hierarchy has a directed acyclic graph structure. The family names are across organims and the family name dictionary is common to each organism. The family database is available from http://marine.ims.u-tokyo.ac.jp:8080/Dict/family. Currently (January, 2004), it contains about 16,000 entries and 70,000 registered names.

4   Gene/protein/family name searches using a devised trie

A gene/protein/family name search of texts was carried out using a devised trie for faster gene name searching. The trie was provided for each organism separately. The core terms implemented for the trie were generated based on GENA. Here, the following main heuristics were used.

(1) Special characters are replaced by a space.

(2) In principle, both numerical and Roman numerals are prepared.

(3) The space before a numerical number is removed. However, if the previous character before the space is a number, the space is not removed (e.g., 14-3-3 is "14 3 3").

(4) With space and without space terms are used for 'Greek alphabet and alphabet a/A, b/B, c/C, ...'. For example, "14 3 3 alpha, 14 3 3alpha, 14 3 3 a, 14 3 3a".

(5) Common words at the end of gene names, such as "protein", "gene", "sub-family", "family", and "group", are removed. However, if the meaning of names is changed with/without these words, they are left. For example, "T-cell surface protein" indicates "protein on the T-cell surface", while "T-cell surface" usually indicates "the surface of the T-cell", and removing "protein" from "memory-related protein" causes faulty recognition of "memory-related function" as 'memory related /gene-name' 'function'. When "protein", "gene", "sub-family", "group", and "family" appear within gene names, gene words with and without these words are generated.

(6) For symbol-type names (less than seven characters), the initial of the organism is added to the spelt-out type. For example, in MAPK1 for *H. sapiens*, hMAPK1 and h MAPK1 are used. For *S. cerevisiae*, the protein name is generated by adding "p" at the end of the name. For example, the protein of STE7 is STE7p. For mutations of *D. melanogaster*, + added names are used. For example, *lt+* for *lt*.

(7) All names are converted into small characters and plurals are also generated. Some names are "case sensitive" and some require "all capital letters". In principle, when the name is the common spelling of a "common noun, adverb, or adjective", "all capital letter names" are adopted in *H. sapiens*, *M. musculus,* and *R. norvegicus* (using "word net vocabularies" with less than five characters. Word length is limited to remove

words that happen to have the same spelling but without removing biological names registered in the word net). "All capital letters names" were recognized in the trie. Case-sensitive words such as cAMP and CAMP were selected experientially and checked after the trie search. Since many of *Drosophila melanogaster* genes have the same spelling with verb, adjective, common nouns, and preposition. These gene names are replaced by "gene name + specified names" using word-net vocabularies to decrease false positive. For example, the gene name "yellow" is replaced by "yellow locus", "yellow gene", "yellow protein", "yellow allele"... etc.

The trie search starts from the next characters after a "space", "-", "/", or "period" or the head of sentence. When multiple gene names are hit in duplicate, the longest name ID is outputted. When specific terms, such as "antagonist", "receptor", "cell", and "inhibitor", ...are next to the gene name, the hit gene name ID is not outputted, since these indicate different gene/protein names or are not gene/protein names. Also, when terms such as "promoter" and "mutant" are located next to the gene name, they do not show the gene/protein/family themselves. However, for our purposes of extracting the genetic interaction, they are treated the same as gene/protein/family names. Specific terms such as "number" are located before the gene name and the hit gene name ID is not outputted since they are multi-sense words and, in most cases, are not gene/protein names. Parentheses are also specially treated, so "mitogen activated protein kinase (MAPK) 1" --> is recognized as "mitogen activated protein kinase 1 (MAPK1)". The continuous gene description such as "GATA-4/5/6" is also specially treated as shown in Figure 1. If the gene names are synonyms of multi-genes, the multiple gene IDs are outputted in this stage.

Figure 1. The schematic drawing of a devised trie.

5 Resolving multi sense words

To resolve the problem of multi-sense words, we used information from the whole text. When the hit name is shorter than a certain gene name length (seven characters for *H. sapiens*; the length is different for each organism), there is a possibility that the hit name is an abbreviation of another word (not only gene names, but also an experimental method or name of an apparatus). To avoid false-positive words as far as possible, we used the following heuristics in *M. musculus, R. norvegicus,* and *H. sapiens*.

1) If the corresponding full name, or a name longer than six characters, is written in the same abstract, the hit gene ID is used.

When the full name and abbreviation pairs are written in the abstract as "plausible full name (the hit name)" or "plausible full name [the hit name]", the following procedures are carried out.

2) If the full/long name is a complete match for the synonyms or full name of the corresponding ID, the hit gene ID is used.

3) If the full/long name is not a complete match for these corresponding IDs using the abbreviation extraction algorithm (Schwartz and Hearst, 2003), but its spelling consists of words used in any name of the corresponding ID, the hit ID is adopted. If not, the hit ID is discarded (i.e., the full/long name considering the replacement of the word order).

4) If information on full names or long names is not found in the abstract, a key-word search of all the abstracts is carried out. If at least one key word is detected, the ID is used.

The summary of these steps were shown in Figure 2. (The numbers in Fig.2 correspond to the above head numbers.)

However, treatment (2) is not sufficient in some cases because some abbreviations are written only once for one family kind. For example, in PUBMED-ID 8248212, ..."the recently described TAP (transporter associated with antigen processing) genes have been mapped approximately midway between DP and DQ. ... In addition to the alleles of TAP1 that have been described, others were identified during this study." "TAP1" is the synonym for "transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)", and "transient receptor potential cation channel, subfamily C, member 4 associated protein." In most cases, the full name is written only once for the same family. In this case, the former ("transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)") is correct. Accordingly, the full name and abbreviation pair "TAP" without the number is also checked. Since all vocabularies ("transporter", "associated", "antigen", "processing") are components of synonyms of TAP1, the TAP1 is recognized by "transporter 1, ATP-binding cassette, and sub-family B (MDR/TAP)". In considering syntactic variations, some

prepositions such as "of" and "with", and frequently used words such as "sub-family" and "family", are skipped in this process. Further regarding the lexico-semantic pattern, as far as possible, adjectives and nouns are provided for each vocabulary using word-net vocabularies and UMLS.
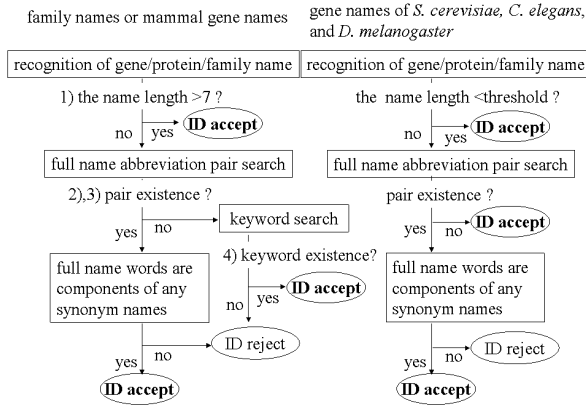

Figure 2. The schematic drawing of each gene names.

With this treatment, only when pairs of full names, or close to the full name, and abbreviations appear, the distinctions between some synonyms are completed. In some cases, the name belongs to the same family. For example, LRE2 is a synonym for "LINE retrotransposable element 2" and "LINE retrotransposable element 3". In this case, the distinction between them is very fine and seems unimportant. In some abstracts, full names are not written in the text. To resolve this issue, we used key words for each gene, which were selected from all words/terms (continuous words) composing synonym names and their family names as shown in the procedures in (4). When at least one keyword is detected, the ID is accepted. The key words appear less than 50 times (only for words extracted from gene names, in the case of words from family name, this limitation is not used) in genes and appear less than a certain frequency in all abstracts and are not common to different genes that have synonyms with the same spelling. Even if a key word search is performed, except for famous names such as p53 and p38, the locus identification for "# kDa", meaning a "#p" expression such as p60 and p61, is quite difficult. In relation to famous name-Ids, such as cAMP(cyclic AMP), CD2(cluster designation 2), the IDs are used to recover a false negative even if the full/longer name is not written in the abstracts and the keywords are not detected.

The automatic keyword selection using conventional methods such as tf-idf (Salton and Yang, 1973) and SMART (Singhal et al. 1996) may be applicable. However, the number of abstracts per gene is too small in many cases and the effective keywords selection could

not be achieved. Therefore, this approach was not applied, in this study.

For *S. cerevisiae, C. elegans,* and *D. melanogaster,* in most cases, the full names of symbols are not written. Only when the symbol name has a symbol (abbreviation)-full name pairs, and the full name is not the corresponding gene name or contains a word that is not a component of the synonyms, the hit-ID is discarded.

Although, as far as possible, we removed what we assumed were wrong or inappropriate gene names, some names either do not seem to be synonyms or are rarely used ones. These can cause errors. For example, LPS is a synonym for "interferon regulatory factor 6" (for example, LocusLink, GenAtlas) and "lipopolysaccharide" in *H. sapiens.* However, our investigations indicate that LPS is not used to indicate "interferon regulatory factor 6" in abstracts.

6 Experiment and Results

To validate the recall and precision of our method for gene/protein/family name recognition, we made manually pre-tagged 100 abstracts (1996 year) on each of the following organisms: *S. cerevisiae*, *D. melanogaster*, and *H. sapiens* with mesh terms "saccharomyces cerevisiae", "drosophila melanogaster", and "human",respectively. Table 1 shows the results. In this evaluation, whether each gene/family ID was correctly assigned in the abstract or not was investigated. (each ID was counted only once per abstract.) When the precision and recall of all gene/family name descriptions' recognition were calculated (each ID can be counted more than once per abstract), they did not change largely and were within 2-5% error spans of Table 1.

Table 1 The summary of precision and recall of gene/protein/family name recognition

Organism*	Precision =TP/(TP+FP) : total(gene/family)	Recall =TP/(TP+FN): total(gene/family)
HS	94.3 (95.2/93.2)%	88.6 (92.0/85.0) %
DM	92.1 (90.3/94.5)%	91.2 (91.8/90.4)%
SC	95.5 (94.6/96.0)%	94.6 (96.0/93.7)%

*HS:*H. sapiens*, DM:*D. melanogaster*, SC:*S. cerevisiae*

The corpus size and the number of deficient name entries in GENA and family name dictionary were summarized in Table 2.

Table 2 The corpus size and num. of deficient gene/family entries.

Organism	Num of gene/family in the corpus: total (gene/family)	Num of deficient name entries: total(gene/family)
HS	167 (87/80)	10 (1/9)
DM	547 (317/230)	31 (16/18)
SC	277 (100/177)	14 (2/11)

In judging family name recognition, slightly soft criteria were used. If a complete matching entry was not registered in the family name dictionary, a higher concept ID was assigned. For example, "lactate dehydrogenase" was not registered in the family name dictionary, so this name was assigned the ID "dehydrogenase". Even if the other organisms are written in the same abstracts, their gene names are not extracted in principle. However, human, rat, and mouse are not distinguished in this validation. The family names in other organisms are also extracted in this evaluation.

As shown in Table 2, in all organisms, more than one-third of the gene names were written as family names. This indicates the necessity for hierarchical gene names, as in the family dictionary, although conventional methods scarcely mentioned. The recall and precision of these organisms as shown in Table 1 are relatively high roughly compared to previous reports. (precision:72-93%, recall:76-94%: The summary is reviewed by Hirschman 2002). The details of errors were as followings. Only 4 and 1 names, which were registered in GENA and family name dictionary, were recognized as gene/family names at once, but they were erroneously discarded by the procedures used to confirm ambiguous names, in *H. sapiens*. Many of them are caused by the key-word search fails. Especially, in family names, the key-words seem to be insufficient. Probably, these will be addressed in some extent by use of the key words of the higher/lower concept IDs. In some cases, the full-name and abbreviation match failed. For example, in "urokinase-type plasminogen activator receptor (uPAR, CD87)", the full-name and abbreviation match failed due to the existence of "two names" in the parenthesis. These errors will be recovered by the keyword search. However, in the present program, recovering step is not used. The recall of family names in *H. sapiens* is slightly low because of varieties of families as shown in Table 1. 6, 4 names were false positive gene/protein names in *S. cerevisiae* and *H. sapiens,* respectively. 7, 5 names were false positive family names in *S. cerevisiae* and *H. sapien,* respectively. Most of

them were short names and were not removed due to their in-appropriate keywords. Some of them are caused by inappropriate GENA entries.

In relation to *D. melanogaster*, 10 gene/protein names that were registered in GENA were not recognized as gene/family names. Many of them were general nouns/adjective and were not used as the "gene name + specified words" phrase in the abstracts. Rest of them were gene/protein names removed in trie implementation steps due to their confusing spellings such as "10-4". Also mutant gene name recognition was quite difficult in this method, since the superscript for the mutation was converted in the normal characters in NCBI-abstracts and newly developed mutant was expressed by changing the superscript. 4 family names were recognized once and erroneously discarded in the keyword search steps. 31 gene/protein names and 12 family names were false positive. Most of them in gene/protein names were misleading names such as 19A. These misleading names were removed or replaced by the "gene name + specified words" phrase as far as possible with some heuristics and term frequencies in abstracts. However, some remained. Some false positive were wrongly extracted other organisms' gene names.

In the strict criteria of family name recognition, 10, 18, 10 names were recognized as higher concepts in H. *sapiens, D. melanogaster,* and *S.cerevisiae,* respectively. The registration of detailed entries for the family name dictionary is required.

The heuristics of the name detection seem to be sufficient so that no name detections failed due to trivial name variations in *H. sapiens and S. cerevisiae,* and only one name in *D. melanog*aster except mutant variation failed. There is some room to be improved in ambiguity resolution steps using sophisticated keyword searching.

In our laboratory, protein interaction information and protein function were automatically extracted and stored in PRIME (http://prime.ontology.ims.u-tokyo.ac.jp) and in the protein kinase database (http://kinasedb.ontology.ims.u-tokyo.ac.jp, Koike et al., 2003). With this procedure, some false positives were not extracted since the phrase patterns did not match the extracted protein interaction and protein function. That is, some wrongly recognized names were removed as a result of considering the local context. In this stage, the wrongly recognized false positive names was 0, 4, and 3 for *S. cerevisiae*, *D. melanogaster,* and *H. sapiens*, respectively. Using the family name dictionary greatly increased the recognition of ambiguous names. However, a new difficulty was found in extracting information. Many family names are common to functional nouns. Therefore, even if a phrase pattern is used, the wrong interaction may be extracted. For example, from PUBMED_11279098: "We also identified key residue

pairs in the hydrophobic core of the Cet1 protomer that support the active site tunnel and stabilize the triphosphatase in vivo." It is difficult to automatically judge from this sentence whether "triphosphatase" means the Cet1 function or another protein family name. All the interaction information in this abstract indicates that "triphosphatase" is the activity of Cet1. Our program wrongly extracted "Cet1/gene-name" stabilize "triphosphatase/family-name". Additional heuristics are required to remove these wrongly extracted data.

7   Related Work

Various protein/gene recognition methods have been reported and some successes were gained as briefly reviewed in introduction and well reviewed in the references (*Hirshman et al.*, 2002). However, most of them did not specify the gene locus. Further, they were developed mainly for *H. sapiens*. Since the naming convention is different in organisms, their recognition performance in other organisms is unknown.

Hirshman et al. (2002) have reported the dictionary-based name recognition. This report discussed the difficulty of the gene name recognition of *D. melanogaster* and showed the increase of the precision by removing the gene names that have meanings as normal English words. Tuason *et al.* (2004) have investigated that the ambiguity within each organism and among organisms (mouse, worm, fly, and yeast) and with general English words. *Tsuruoka and Tsujii* (2003) also reported the dictionary-based named recognition and our method is similar to them. They resolved the trivial gene variation problems using dynamic programming and tries, while in our method, by normalizing dictionary names and devising the trie structure, the trivial variations were addressed without dynamic programming and the required CPU time is expected to be largely reduced without decreasing precision and recall. The protein name recognition standard is a little different from them and the direct comparison of precision and recall with their results seem meaningless. In their methods, they focus on protein names (without gene names) and seem not to distinguish whether the protein name candidate represents the protein itself or not in the context. (ex. "IL-1 receptor antagonist" and "IL-1 receptor expression": only the latter description means the IL-1 receptor itself.)   Further, in our method, addressing the ambiguity of gene names (common gene names among multiple gene names) is tried. Since long protein names are usually written with abbreviated names, the name variations caused of permutation and insertion/deletion of long name words are picked up in the ambiguity resolution process.

8   Conclusions:

We constructed gene name and family name dictionaries to link each gene name to a gene locus and to relate ambiguous names to gene families. Our preliminary investigations showed that more than one-third to one-half of gene/protein names in abstracts are written using ambiguous names such as family/super-family level names. This indicates that dictionary-based gene/protein/family name recognition requires not only a gene name dictionary but also a hierarchical family name dictionary. Using the gene name dictionary GENA and the family name dictionary we constructed and our searching method, 95, 91, and 89% of protein/gene/family names in abstracts on *S. cerevisiae, D. melanogaster,* and *H. sapien*s were detected with a precision of 96, 92, and 94%, respectively. The simple heuristics we developed seem to be useful for matching gene/family names in texts with dictionary entry names, although additional trivial changes are required to address ambiguity of gene names. These methods are also useful for extracting data on protein interaction and protein function. However, the gene/protein/family name recognition subject is deep. For example, "NFkappaB" represents "NFKB1" and "RELA" complex in many contexts and sometimes represents "NFKB1". Unfortunately, these complicated recognitions were not resolved.

Although different organisms have different naming conventions, the nomenclature for mammals is similar to that for *H. sapiens,* and most bacteria and archaea gene/protein/family names are similar to the nomenclature for *S. cerevisiae*. Problems in gene name recognition for most organisms will be able to be addressed using our method. Dictionary-based name recognition cannot search new gene name/synonym names. However, the whole human/drosophila/yeast genomes have already been sequenced and the appearance of new synonym names can be expected to decrease or be inferable from the referenced known name. In addition, with the introduction of the family name dictionary, parts of new genes can be retrieved using the higher concept name (family name), even if the new gene name itself is not registered in GENA. Accordingly, the dictionary-based name recognition will be expected to be sufficient for the information extraction in these organisms.

Protein-interaction and protein-function information extracted using these procedures for gene/protein/family name recognition are available from http://prime.ontlogy.ims.u-toky.ac.jp.

Acknowledgements

References

Collier, N., Nobata, C. and Tsujii, J. 2000. *Proc. of the 18th Int. Conf. on Comp. Ling.* 201-207.

Enright, AJ, Van Dongen, S, and Ouzounis, CA. 2002. *Nucleic Acids Res*. 30(7):1575-84.l

Fukuda, K., Tsunoda, T. Tamura, A. and Takagi, *T. 1998 Proceedings of the Pacific Symposium on Bio-computing,* 705-716.

Hatzivassiloglou, V., Duboue, P.A. and Rzhetsky, A. 2001. *Bioinformatics*, 17S(1), S97-S106.

Hirschman, L., Morgan, A.A., and Yeh, A.S. 2002 *J. Biomed. Inform*. 35:247-259.

Humphreys, K., Demetriou, G., and Gaizauskas, R. 2000 *Proc. of the Pacific Symposium on Biocomputing,* 5:502-513.

Jenssen TK, Laegreid A, Komorowski J, Hovig E. 2001. Nat Genet. 28(1):21-8.

Koike, A., Kobayashi, Y., and Takagi, T. 2003. *Genome Research*, 13:1231-1243.

Nobata, C., Collier, N., and Tsujii, J. 1999. *Proc. of Nat. Lang. Paci. Rim Symp*. 369-374.

Salton, G. and Yang, C.S. (1973) J. Document. 29(4), 351-372.

Schwartz, A.S., Hearst M.A., 2003. *Pacific Symposium on Biocomputing* 8:451-462.

Singhal, A. Buckley, C., and Cochrane, P.A. 1996. *Proc. of ACM SIGIR*, 26-133.

Tanabae, L and Wilbur,WJ. 2002. *Bioinformatics*, 18(8):1124-1132.

Tsuruoka, Y. and Tsujii, J. 2004. *Proc. of the ACL 2003 Workshop on Natural Language Processing in Biomedicine* 41-48.

Tuason, O. and Chen, L., Liu, H., Blake, J.A., and Friedman, C. 2004. *Proc. of Pacific Symposium on Biocomputing*, 238-249.