Vol. 16 no. 2 2000
Pages 169–175

PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary

Mikio Yoshida, Ken-ichiro Fukuda and Toshihisa Takagi

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo, 108-8639, Japan

Received on April 1, 1999; revised on August 18, 1999; accepted on November 30, 1999

Abstract

Motivation: *Since their initial development, integration and construction of databases for molecular-level data have progressed. Though biological molecules are related to each other and form a complex system, the information is stored in the vast archives of the literature or in diverse databases. There is no unified naming convention for biological object, and biological terms may be ambiguous or polysemic. This makes the integration and interaction of databases difficult. In order to eliminate these problems, machine-readable natural language resources appear to be quite promising. We have developed a workbench for protein name abbreviation dictionary (PNAD) building.*

Results: *We have developed PNAD Construction Support System (PNAD-CSS), which offers various convenient facilities to decrease the construction costs of a protein name abbreviation dictionary of which entries are collected from abstracts in biomedical papers. The system allows the users to concentrate on higher level interpretation by removing some troublesome tasks, e.g. management of abstracts, extracting protein names and their abbreviations, and so on. To extract a pair of protein names and abbreviations, we have developed a hybrid system composed of the PROPER System and the PNAD System. The PNAD System can extract the pairs from parenthetical-paraphrases involved in protein names, the PROPER System identified these pairs, with 98.95% precision, 95.56% recall and 97.58% complete precision.*

Availability: *PROPER System is freely available from http://www.hgc.ims.u-tokyo.ac.jp/service/tooldoc/KeX/intro.html. The other software are also available on request. Contact the authors.*

Contact: *mikio@ims.u-tokyo.ac.jp*

Introduction

Stimulated by the Human Genome Project, integration and construction of databases for molecular-level data, which are rather isolated and non-hierarchical in contrast to the situation in the living cell, have progressed significantly. However, as biological molecules are related to each other and form a complex system (e.g. the living cell), interpretation of the genome requires information about relations such as cell-level information (e.g. protein interaction data) and individual-level information (e.g. the relation between disease and mutations). These information would be acquired from the literature and by integrating and reorganizing a multiplicity of molecular-level data into biological units.

Unfortunately, these tasks are not trivial. First, extracting information from the literature to annotate database entries is a time-consuming task and requires curatorial expertise in both the text retrieving phase and the information extraction phase. Therefore, systems that reduce this labor are needed (Andrade and Valencia, 1998; Ohta *et al.*, 1997). Natural language processing techniques would be very helpful, but they rely on the quality of lexicons. Secondly, there is no unified naming convention of biological objects (Fukuda *et al.*, 1998) and, even worse, the meaning or the specification of biological concepts may differ between databases [e.g. the meaning of the term 'gene'(Schulze-Kremer, 1998)]. The same term may have different meanings, and different terms may have the same meaning. This makes the integration of databases non-trivial.

Preparation of machine-readable natural language resources (e.g. synonym dictionaries and abbreviation dictionaries for genes and proteins) is one thing to do to overcome these difficulties.

For example, styles of abbreviations can take various patterns as shown below. And the mapping from the original word to the abbreviation is not one-to-one, which is a nuisance for both database curators and end users. Here we show six styles of abbreviation:

type 1. consists of initial characters
Thyrotrophin-releasing hormone (TRH)

type 2. consists of capital or numerical characters
IL-2 receptor sub-unit (IL-2R)

type 3. consists of initial characters of syllables
Inter·cel·lu·lar ad·he·sion mole·cule 1 (ICAM1)
ni·tro·glyc·er·in (NTG)

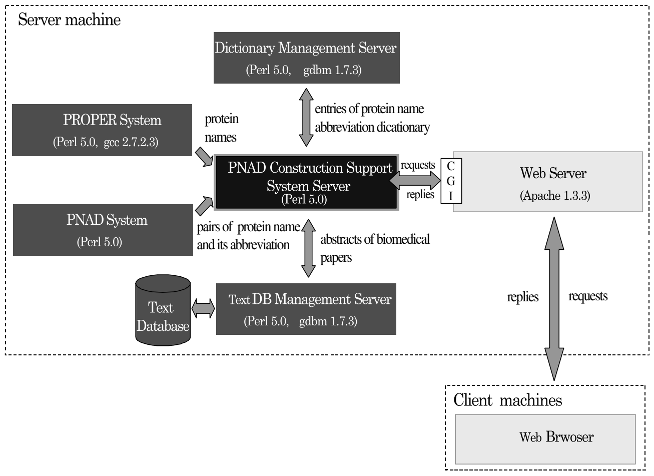


Fig. 1. System architecture of the PNAD-CSS. The sub-systems are depicted as squares, and arrows between sub-systems indicate communications.

type 4. consists of some characters of which a full name is composed
Meges·trol ac·e·tate (megace)

type 5. the order of some characters in an abbreviation are changed (inversion)
NG-monomethyl-L-arginine (L-NMMA)

type 6. Some characters are substituted by other expressions
fi·bro·nec·tin type III (FN3)

where the center dots ('·') separate each term into syllables.

We developed a system that supports the construction of abbreviation dictionaries, which is a step toward the goal.

In the following sections, the system architecture and the method to identify abbreviations from texts are described. After that we show some examples of the way the user can manipulate the system and then discuss applications and future work.

System

The system, called PNAD Construction Support System (PNAD-CSS), stores MEDLINE abstracts (NLM, 1998), tries to identify abbreviations and their corresponding original phrases, offers a management system for dictionary construction, and enables multiple operators to work in an interactive way.

The system is a server client system based on the World Wide Web (WWW) architecture (Figure 1). The server consists of six subsystems; Web Server, PNAD-CSS Server, Dictionary Management Server, Text Database Management Server, PROPER System (Fukuda *et al.*, 1998), and Abbreviation Extraction System. The client is a Web Browser. The server system runs on a Sun workstation under Solaris (Version 2.6). All the programs we developed are implemented in Perl 5 (Wall *et al.*, 1996). The Text database and Dictionary are constructed using gdbm 1.7.3 (library for database functions).

- The PNAD-CSS Server is the core of this system, which executes and controls other programs and servers in response to commands passed from clients via Common Gateway Interface (CGI).
- Currently the system is designed to, but not limited to, the use of MEDLINE abstracts. Considering the availability and the existence of a unified format, the authors decided to use MEDLINE abstracts rather than on-line full papers. As the number of abstracts can reach tens of thousands per month, the system offers a Text Database Management System so that operators can know which abstracts are finished and which abstracts are currently opened by other operators.
- The Dictionary Management Server offers five essential functions to operate a protein name abbreviation dictionary (i.e. to register a new entry, to modify an existing one, to search entries, to delete an existing one, and to merge two existing dictionaries). Each entry of the dictionary has eleven items, a protein name, its abbreviation, a title of a paper, the names of the authors, the name of the journal, the publication date, the species, an example sentence, descriptions about the protein, any worker's comments, and other comments, of which the underlined six items are automatically extracted from MEDLINE abstracts.
- Both the PROPER System and the PNAD System are modules for extracting pairs consisting of a protein name and its abbreviation. The PROPER System tags protein names in the biomedical text, and the PNAD System extracts pairs of an abbreviation and its associated term from *parenthetical-paraphrases* (e.g. 'C-X-C chemokin cytokine induced neutrophil chemoattractant (CINC)', 'large subunit of eukaryotic initiation factor 2 (seIF)', and so on). The brief explanation of the PROPER System and detail of the PNAD System will be described in the next section.

Method

In this section, both the concept of the protein name identification method and the algorithm for identifying pairs of a protein name and its abbreviation are described. Figure 2 shows the overall architecture of the PNAD System. In Phase (1), we used the PROPER System.

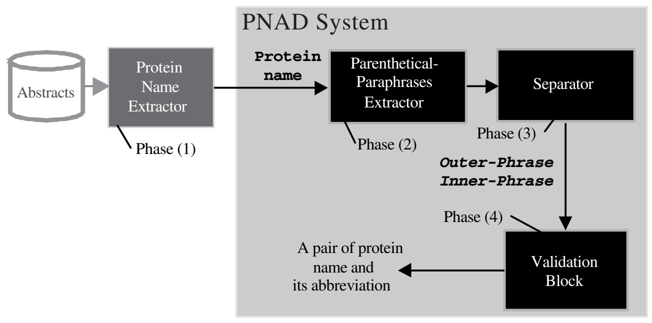
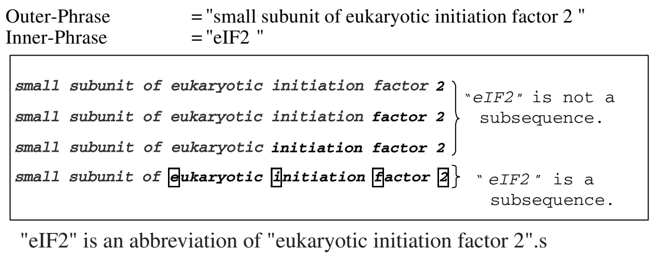


Fig. 2. System architecture of the PNAD System. Separator is for separating a *Parenthetical-Paraphrase* into an *Outer-Phrase* and an *Inner-Phrase* (in this article, we will call the phrase enclosed in bracket *Inner-Phrase*, and the followed by bracket *Outer-Phrase* in a expression of *Parenthetical-Paraphrase*).



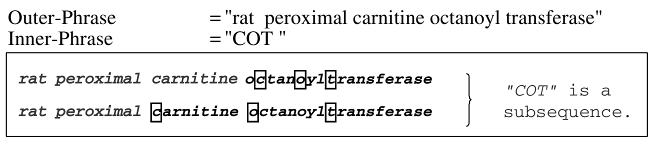
Outer-Phrase = "small subunit of eukaryotic initiation factor 2 "
Inner-Phrase = "eIF2 "

small subunit of eukaryotic initiation factor 2
small subunit of eukaryotic initiation ***factor 2***
small subunit of eukaryotic ***initiation factor 2***
"*eIF2*" is not a subsequence.

small subunit of ***eukaryotic initiation factor 2***
"*eIF2*" is a subsequence.

"eIF2" is an abbreviation of "eukaryotic initiation factor 2".s

Fig. 3. An example of simple validation.



Outer-Phrase = "rat peroximal carnitine octanoyl transferase"
Inner-Phrase = "COT "

rat peroximal carnitine ***octanoyltransferase***
rat peroximal ***carnitine octanoyltransferase***
"*COT*" is a subsequence.

Fig. 4. An example of existing two subsequence candidates.

Protein Name Identification Method

Name identification is not a trivial task, especially in medical/biological documents. One will encounter the following difficulties: unknown word recognition, long compound word recognition, and the requirement of robustness against diversity of expressions for each protein. A typical method used to identify technical terms is to match each word to the heading words on prepared dictionaries. But we do not have those dictionaries and even worse it is not guaranteed that people use the exact spelling defined in the dictionaries. There are also many works on compound word recognition using statistical methods (Su *et al.*, 1994; Smadja, 1993), but medical/biological compound words may include more than six words for which getting discriminative statistical thresholds is difficult.

In PROPER, the authors decided to use a simple rule-based method that uses morphological clues on character strings in medical and biological documents. It utilizes the characteristics of proper noun description in these fields and does not require any specific term dictionary prepared in advance. PROPER was tested on 80 MEDLINE abstracts with the result of 87.20–94.70% precision and 93.32–99.85% recall.

Validation Method

In Figure 3, a box marks the letters that best fit the abbreviation for 'eIF2'. Similarly in Figure 4 for 'COT'. We define the phase that finds these boxes as 'Validation Block' (Figure 2). The 'Validation Block' determines whether the *Inner-Phrase* is an abbreviation of its original term. See Figure 3, 'eIF2' is a subsequence[†] of 'eukaryotic initiation factor 2', so 'eIF2' would be regarded as an abbreviation of 'eukaryotic initiation factor 2'. On the other hand, in Figure 4, although 'COT' is a subsequence of 'octanoyltransferase', it is obvious that 'COT' is not an abbreviation of 'octanoyltransferase'. This suggests that even if a string A is a subsequence of a string B, A is not always an abbreviation of B. The authors assumed that there would be particular characters which tend to compose an abbreviation, for instance initials of words or syllables. Therefore, in the proposed algorithm, the validation block generates a subsequence C from B using characters that tend to be a member of an abbreviation. Then it checks if A is a subsequence of C or not. This would be similar to the way we recognize pairs of an abbreviation and its original term.

Figure 5 represents a block diagram of the following algorithm.

Step (0) Let variable number n be 0.

Step (1) Let variable number m be 0.

Step (2) If m is greater than the number of words composing the *Outer-Phrase*, go to Step (8), otherwise let m be $m + 1$.

Step (3) Extract m words from the end of the *Outer-Phrase*.

Step (4) Make a subsequence S1 from the initial characters and their following n characters of the extracted m words in Step (3).

Step (5) If the *Inner-Phrase* is a subsequence of S1, register the *Inner-Phrase* as an abbreviation and the extracted m words in Step (3) as the original term, then quit this process.

[†]A subsequence is the letters from a string but the letters do not have to be contiguous (Baeza-Yates, 1992).

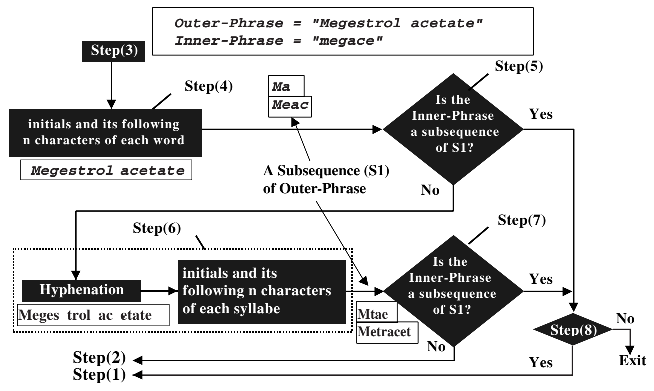

Outer-Phrase = "Megestrol acetate"
Inner-Phrase = "megace"
Step(3)
Step(4)
Ma
Meac
Step(5)
initials and its following n characters of each word
Megestrol acetate
Is the Inner-Phrase a subsequence of S1?
Yes
A Subsequence (S1) of Outer-Phrase
No
Step(6)
Step(7)
Hyphenation
initials and its following n characters of each syllabe
Meges trol ac etate
Is the Inner-Phrase a subsequence of S1?
Yes
Mtae
Metracet
No
Step(8)
No
Exit
Yes
Step(2)
Step(1)

Fig. 5. A block diagram of the validation method.

Step (6) Make a subsequence S1 from the initial words and their following *n* characters of the syllables which are obtained by hyphenating *m* words extracted in Step (3).

Step (7) If the *Inner-Phrase* is a subsequence of S1, register the *Inner-Phrase* as an abbreviation and the extracted *m* words in step (3) as the original term, then quit this process. Otherwise go to Step (2).

Step (8) If the maximum number of characters composing each syllable previously obtained is greater than *n*, let *n* be $n + 1$ and go to Step (1). Otherwise, exit this process.

To hyphenate words, we adopted a hyphenation method which had been used in TEX82 (Knuth, 1984).

Evaluation

We evaluated the proposed method using: (1) the abstracts of 17 173 biomedical papers published in March 1996 (MAR), which includes 17 666 *Parenthetical-Paraphrases*, and a size of 33.9 Mb; (2) the abstracts of 6296 biomedical papers published in July 1996 (JUL), which includes 7295 *Parenthetical-Paraphrases*, and a size of 13.3 Mb. Both were acquired from MEDLINE.

The aim of this evaluation is to estimate the ability of the PNAD System, although the accuracy of the PNAD System largely depends on the ability of the PROPER System. In order to eliminate its influence, we assumed that the PROPER System can extract all of the protein names and that they are completely correct. If the PROPER System extracted an incomplete abbreviation for a *Parenthetical-Paraphrase* (e.g. because of an incomplete original term) and the PNAD System extracted it, we considered it an abbreviation extraction error (an Incorrect Abbreviation). On the other hand, when the PROPER System extracted a term which did not represent a protein name and the PNAD System extracted a correct abbreviation from it, we considered it to be a correct extraction. In addition, we assumed that all protein name abbreviation defining phrases are extracted by the PROPER System.

Table 1. The evaluation results. The boundary failures means that the extracted original term lacks words or has extra words while the extracted abbreviation is correct. The number of the abbreviations categorized in Boundary Failure is involved in that of Correct Abbreviations. Processing Time is that for extracting from the results of PROPER System on a Sun Ultra Enterprise 4000 (Sun OS 5.5.1, CPU Clock 168 MHz, Memory Size 1.5 GB, Virtual Memory Size 3.0 GB)

	Items		Values
(1)	Extracted	Correct Abbrev.	13 567
(2)		(Boundary Failure)	(188)
(3)		Incorrect Abbrev.	144
(4)	Unextracted	Correct Abbrev.	629
(5)		Incorrect Abbrev.	10 620
(6)	Precision	(1)/((1)+(3))	98.95%
(7)	Recall	(1)/((1)+(4))	95.56%
(8)	Complete Precision	((1)–(2))/((1)+(3))	97.58%
(9)	Processing Time (s)		4280

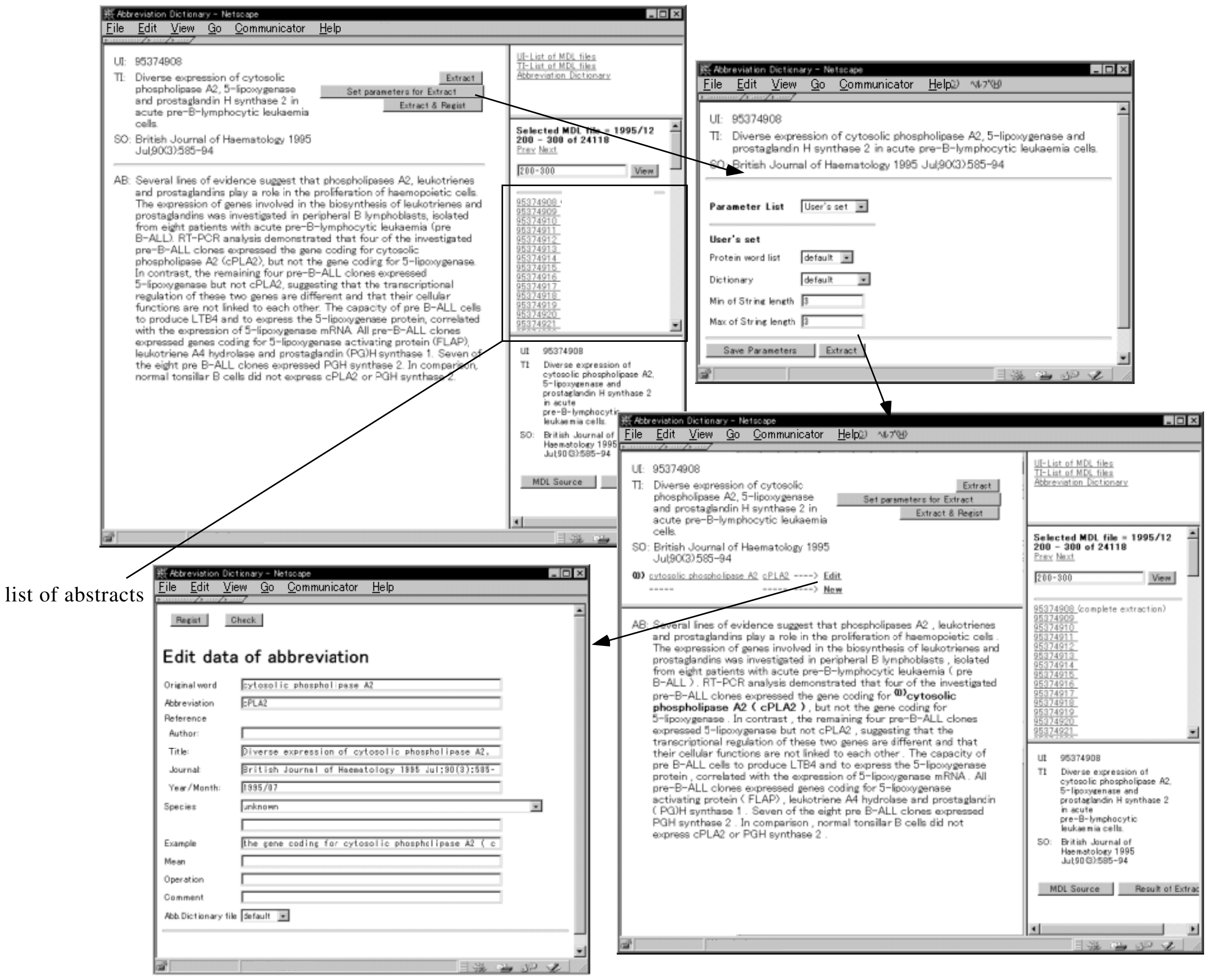
The confirmation of correct abbreviations was determined manually by a non-biomedical specialist. In this evaluation, we considered the terms that the PROPER System extracted were all protein names as mentioned above, therefore, the task did not need any high-level biological knowledge (e.g. to identify correct protein names).

Table 1 shows the abbreviation extraction results using the evaluated data from the abstract sets. ‘Precision’ and ‘Recall’ are popular metrics used in the Information Retrieval (IR) domain, and would correspond with ‘Specificity’ and ’Sensitivity’ respectively in the biological domain. In the ‘Precision’ metric, incompletely extracted abbreviations are also considered to be correct abbreviations (see Table 1). The reason is that, in data entry support systems like this, showing abbreviations would be sufficiently helpful for operators even if they are not completely correct. The ‘Complete Precision’ metric in Table 1 indicates a precision based on completely extracted abbreviations.

Our proposed method extracted abbreviations with high accuracy. This confirms that the validation scheme based on the empirical rules is effective in this case.

Examples of usage

There are two major functions of PNAD-CSS i.e. functions to register or edit entries and to search entries.



UI: 95374908
TI: Diverse expression of cytosolic phospholipase A2, 5-lipoxygenase and prostaglandin H synthase 2 in acute pre-B-lymphocytic leukaemia cells.
SO: British Journal of Haematology 1995 Jul;90(3):585-94

AB: Several lines of evidence suggest that phospholipases A2, leukotrienes and prostaglandins play a role in the proliferation of haemopoietic cells. The expression of genes involved in the biosynthesis of leukotrienes and prostaglandins was investigated in peripheral B lymphoblasts, isolated from eight patients with acute pre-B-lymphocytic leukaemia (pre B-ALL). RT-PCR analysis demonstrated that four of the investigated pre-B-ALL clones expressed the gene coding for cytosolic phospholipase A2 (cPLA2), but not the gene coding for 5-lipoxygenase. In contrast, the remaining four pre-B-ALL clones expressed 5-lipoxygenase but not cPLA2, suggesting that the transcriptional regulation of these two genes are different and that their cellular functions are not linked to each other. The capacity of pre B-ALL cells to produce LTB4 and to express the 5-lipoxygenase protein, correlated with the expression of 5-lipoxygenase mRNA. All pre-B-ALL clones expressed genes coding for 5-lipoxygenase activating protein (FLAP), leukotriene A4 hydrolase and prostaglandin (PG)H synthase 1. Seven of the eight pre B-ALL clones expressed PGH synthase 2. In comparison, normal tonsillar B cells did not express cPLA2 or PGH synthase 2.

Selected MDL file = 1995/12
200 - 300 of 24118

200-300

95374908 (complete extraction)
95374909
95374910
95374911
95374912
95374913
95374914
95374915
95374916
95374917
95374918
95374919
95374920
95374921

UI 95374908
TI Diverse expression of cytosolic phospholipase A2, 5-lipoxygenase and prostaglandin H synthase 2 in acute pre-B-lymphocytic leukaemia cells.
SO: British Journal of Haematology 1995 Jul;90(3):585-94

Parameter List User's set

User's set

Protein word list	default
Dictionary	default
Min of String length	3
Max of String length	3

(0) cytosolic phospholipase A2 cPLA2 ----> Edit
----- ----- ----> New

AB: Several lines of evidence suggest that phospholipases A2 , leukotrienes and prostaglandins play a role in the proliferation of haemopoietic cells . The expression of genes involved in the biosynthesis of leukotrienes and prostaglandins was investigated in peripheral B lymphoblasts , isolated from eight patients with acute pre-B-lymphocytic leukaemia (pre B-ALL) . RT-PCR analysis demonstrated that four of the investigated pre-B-ALL clones expressed the gene coding for (0)**cytosolic phospholipase A2 (cPLA2)** , but not the gene coding for 5-lipoxygenase . In contrast , the remaining four pre-B-ALL clones expressed 5-lipoxygenase but not cPLA2 , suggesting that the transcriptional regulation of these two genes are different and that their cellular functions are not linked to each other . The capacity of pre B-ALL cells to produce LTB4 and to express the 5-lipoxygenase protein , correlated with the expression of 5-lipoxygenase mRNA . All pre-B-ALL clones expressed genes coding for 5-lipoxygenase activating protein (FLAP) , leukotriene A4 hydrolase and prostaglandin (PG)H synthase 1 . Seven of the eight pre B-ALL clones expressed PGH synthase 2 . In comparison , normal tonsillar B cells did not express cPLA2 or PGH synthase 2 .

Edit data of abbreviation

Original word	cytosolic phospholipase A2
Abbreviation	cPLA2
Reference	
Author:	
Title:	Diverse expression of cytosolic phospholipase A2,
Journal:	British Journal of Haematology 1995 Jul;90(3):585-
Year/Month:	1995/07
Species	unknown
Example	the gene coding for cytosolic phospholipase A2 (c
Mean	
Operation	
Comment	
Abb.Dictionary file	default

Fig. 6. An example of a procedure for registration. Target abstract is shown in the top-left window. The top-right window is for setting extraction parameters. In the bottom-right window, the result of the extraction is presented. The bottom-left window is for editing and registration of an entry.

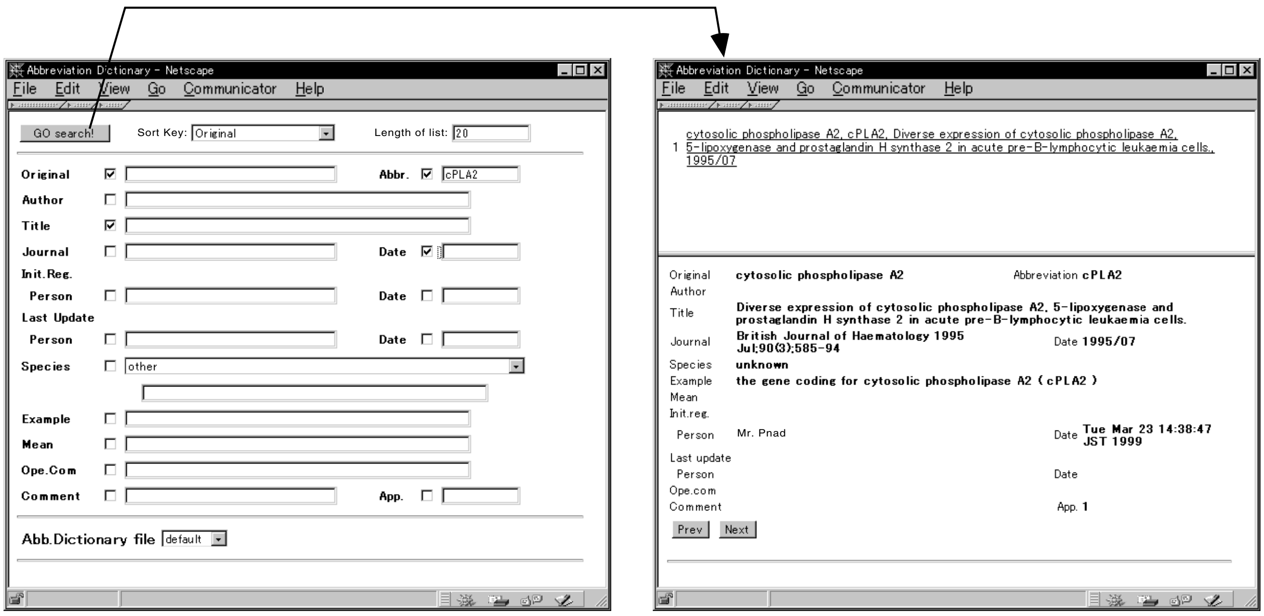
Registration

There are five steps to register entries into the dictionary: selecting an abstract, executing abbreviation extraction from it, choosing an extracted abbreviation, completing the entry of the dictionary, and registering it. Figure 6 illustrates the procedure.

The first step is to select an abstract. The users can select one from the lists shown in the right-middle frame of the top-left window. In the list, each abstract is represented by its MEDLINE UI (e.g. 95374908), and '`(complete extraction)`' following this UI number indicates that abbreviation extraction from that abstract has been carried out. By clicking any MEDLINE UI, the title and journal name of each abstract is shown in the right-bottom frame, and full text of the abstract is displayed in the left large frame.

The next is to extract protein name and abbreviation pairs which is done automatically by the PNAD System. Prior to this, users can set parameters for the PNAD System (e.g. the minimum length of abbreviations, hyphenation dictionary name, and so on) as shown in the top-right window.

After extraction has been carried out, users can see the results. In the bottom-right window of Figure 6, the extracted protein name and its abbreviation are shown at the bottom of the left-top frame, and in the left-bottom frame, users can see where the associated phrase is located in the abstract. By clicking a word '`Edit`' following the result in the left-top frame, a window for editing the entry will appear (bottom-left window). As mentioned above, the protein name, its abbreviation, journal title and its published date are filled in automatically, so users need not


Abbreviation Dictionary - Netscape
File Edit View Go Communicator Help
GO search!
Prev
Next

Sort Key: Original
Length of list: 20

Original ☑
Abbr. ☑ cPLA2
Author ☐
Title ☑
Journal ☐
Date ☑
Init.Reg.
Person ☐
Date ☐
Last Update
Person ☐
Date ☐
Species ☐ other
Example ☐
Mean ☐
Ope.Com ☐
Comment ☐
App. ☐

Abb.Dictionary file default

1 cytosolic phospholipase A2, cPLA2, Diverse expression of cytosolic phospholipase A2, 5-lipoxygenase and prostaglandin H synthase 2 in acute pre-B-lymphocytic leukaemia cells., 1995/07

Original	cytosolic phospholipase A2	Abbreviation	cPLA2
Author			
Title	Diverse expression of cytosolic phospholipase A2, 5-lipoxygenase and prostaglandin H synthase 2 in acute pre-B-lymphocytic leukaemia cells.		
Journal	British Journal of Haematology 1995 Jul;90(3);585-94	Date	1995/07
Species	unknown		
Example	the gene coding for cytosolic phospholipase A2 (cPLA2)		
Mean			
Init.reg.			
Person	Mr. Pnad	Date	Tue Mar 23 14:38:47 JST 1999
Last update			
Person		Date	
Ope.com			
Comment		App.	1

Fig. 7. An example of a procedure for search. The left is a search window, and the right is a result window.

fill in these fields. The entry will be registered by clicking the ‘`Regist`’ button on top of the window.

Search

The system has several convenient functions to search existing entries in the dictionary. The procedure of searching entries is shown in Figure 7. The left window is for setting various parameters for search and the right one is for showing the result of the search.

Discussion and future work

The proposed system contributes towards construction of a protein name abbreviation dictionary by automatically extracting candidates of abbreviations and their associated original names.

While automatic methods cannot be 100% correct, manual methods are time-consuming and may generate artifacts (e.g. typo’s). In this system, operators construct an abbreviation dictionary in an interactive way with the support of automatic annotation systems. The system will compensate for the draw-backs of both automatic and manual methods and reduce the cost of such dictionary construction procedures.

However, the system has two restrictions. First, the algorithm in the Validation subsection will not recognize some types of abbreviation, defining phrases as correct. Those are type 5 and 6 in the introduction (for example, ‘L-NMMA’ for ‘NG-monomethyl-L-arginine’ and ‘IL-11’ for ‘interleukin eleven’). We examined the number of each type of *Parenthetical-Paraphrases* in abstracts of biomedical papers which are published in March 1996. The number of abbreviation defining *Parenthetical Paraphrases* was 7962, where 6096 of them (97%) were classified into Type 1–4. This indicates that the type of abbreviation defining phrases that our system cannot recognize is relatively rare. We expect those abbreviations will appear in a valid form in other abstracts and will eventually be registered in the dictionary if a sufficient number of abstracts are processed.

Second, the current system does not utilize its output (i.e. its abbreviation dictionary) to identify names in other abstracts owing to the module PROPER. PROPER is a kind of *ab initio* prediction tool and was designed not to use biomedical dictionaries. As an abbreviation dictionary will now become available, feeding back the output into the system will reduce processing time and annotation errors.

Besides the lack of dictionaries, the considerable variety of nomenclature rules or guidelines among each model organisms (Reid, 1998) will be an obstacle to the construction of an overall system that supports all model organisms. For this reason, we constructed a support system for dictionary construction that does not require other dictionaries. This effort will contribute to future research (developments) in this area.

Acknowledgements

We thank members of the Human Genome Center who advised us. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, ‘Genome

Science', from the Ministry of Education, Science, Sports and Culture in Japan, and JSPS Research Project for the Future.

References

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

Baeza-Yates,R.A. (1992) Information retrieval data structures & algorithms. In Frakes,W.B. and Baeza-Yates,R.A. (eds), *Introduction to Data Structures and Algorithms Related to Information Retrieval.* Prentice Hall, New Jersey, pp. 13–27.

Fukuda,K., Tamura,A., Tsunoda,T. and Takagi,T. (1998) Toward information extraction identifying protein names from biological, papers. In *Proceedings of the Pacific Symposium on Biocomputing '98,* pp. 707–718.

Knuth,D.E. (1984) *The TEXbook*. Addison-Wesley, Massachusetts.

NLM, (1998) *PubMed NLM's search searvice*. http://www.ncbi.nlm.nih.gov/PubMed/. National Library of Medicine.

Ohta,Y., Yamamoto,Y., Okazaki,T., Uchiyama,I. and Takagi,T. (1997) Automatic construction of knowledge base from biological papers. In *Proceedings of the ISMB–97*, **5**, 218–225.

Reid,T. (1998) In Wood,R. (ed.), *Trend in GENETICS: Genetic Nomenclature Guide* Elsevier Trends Journals, Cambridge.

Schulze-Kremer,S. (1998) Ontologies for molecular biology. In *Proceedings of the Pacific Sypmosium on Biocomputing '98,* pp. 695–706.

Smadja,F. (1993) Retrieving collocations from text: xtract. *Computational Linguistics*, **19** (1).

Su,K., Wu,M. and Chang,J. (1994) A corpus-based approach to automatics compound extraction. In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94).*

Wall,L., Christiansen,T. and Schwartz,R. (1996) *Programming Perl*. 2nd edn, O'Reilly & Associates.