# A Keyword Recommendation System for GenBank

**Yasuhiko Kitamura**
kitamura@info.eng.osaka-cu.ac.jp

**Tetsuro Nanbu**
tetsuro@kdel.info.eng.osaka-cu.ac.jp

**Shoji Tatsumi**
tatsumi@info.eng.osaka-cu.ac.jp

Department of Information and Communication Engineering, Osaka City University,
3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 558-8585, Japan

## 1 Introduction

With the progress of Human Genome Project, a huge amount of genomic data is stored in various databases and is open to the public through Internet. The size of these databases grows rapidly and, for example, the size of GenBank[1], which is a well-known DNA sequence database, reaches more than 5 million entries. However, as the size of database grows, it becomes more difficult for users to retrieve necessary data. Many of genome databases consist of full text data and the keyword search method is widely used for their retrieval. Hence, when a user submits a simple query with a general keyword, he receives a large number of search results and is obliged to add another keywords to reduce the number of results.

In this paper, we propose a keyword recommendation system [1, 2] that recommends proper keywords to the user and show a prototype where the technique is applied for GenBank. Advantages of our system are as follows.

- It assists users who have difficulty to think of proper keywords to narrow down the search.
- It may lead to creative thinking support for researchers by showing keywords related to their original inputs.

## 2 Method

To recommend proper keywords, we use keyword pairs of co-occurrence as the candidates. A keyword pair of co-occurrence is a pair of keywords that appear in an identical entry. For example, in "mouse liver milia Mus musculus cDNA clone," "mouse" and "cDNA" are a keyword pair of co-occurrence. Then, for "mouse" as an original input, "cDNA" can be a keyword candidate for recommendation. We sort co-occurred keywords according to the frequency.

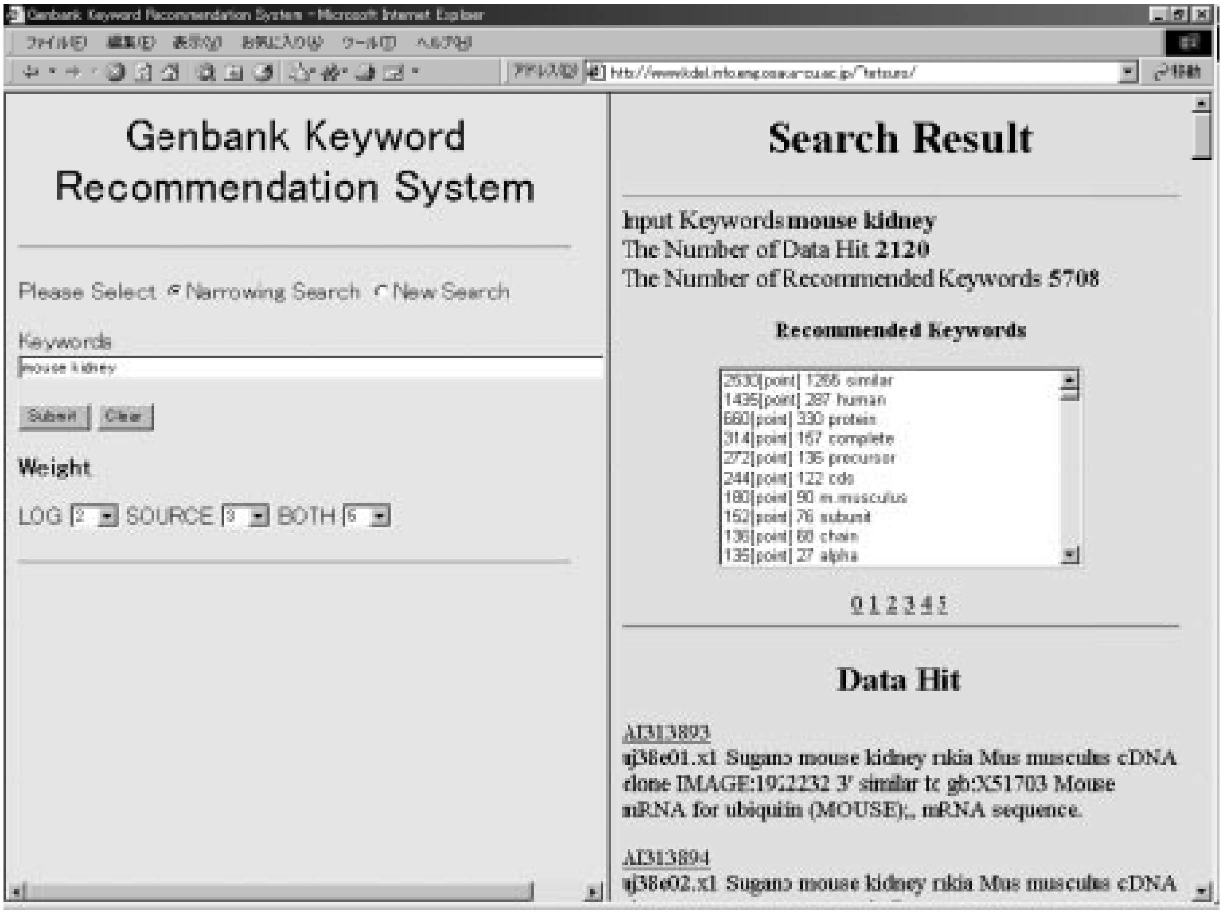
In addition, we can use thesaurus and log data to sort the keyword more elaborately. Thesaurus is a kind of dictionary that contains biological terms and can be used as a measure that shows the significance of keywords in general sense. Log data is a record of input keywords and can be used as a measure that shows keyword preference of a user or a group of users.

## 3 Prototype

We developed a prototype for GenBank[2] as shown in Fig. 1. For testing, we use only 80,000 entries of GenBank at this moment. At first, we start to use this system by giving one or more query keywords

[1] http://www.genome.ad.jp/dbget-bin/www_bfind?genbank-today

[2] http://www.kdel.info.eng.osaka-cu.ac.jp/~tetsuro/


Genbank Keyword Recommendation System - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D)
Submit
Clear

Genbank Keyword Recommendation System

Please Select ● Narrowing Search ○ New Search

Keywords: mouse kidney

Weight

LOG 2 SOURCE 3 BOTH 6

Search Result

Input Keywords **mouse kidney**
The Number of Data Hit **2120**
The Number of Recommended Keywords **5708**

**Recommended Keywords**

2530[point] 1265 similar
1435[point] 287 human
660[point] 330 protein
314[point] 157 complete
272[point] 136 precursor
244[point] 122 cds
180[point] 90 m.musculus
152[point] 76 subunit
136[point] 68 chain
135[point] 27 alpha

0 1 2 3 4 5

Data Hit

AI313893
uj38e01.x1 Sugano mouse kidney mkia Mus musculus cDNA clone IMAGE:1922232 3' similar to gb:X51703 Mouse mRNA for ubiquitin (MOUSE);, mRNA sequence.

AI313894
uj38e02.x1 Sugano mouse kidney mkia Mus musculus cDNA

Figure 1: GenBank Keyword Recommendation System.

as usual in the left frame. By clicking "Submit," the systems returns recommended keywords, links to data entries, and other statistics in the right frame. By clicking a recommended keyword in the box, the system automatically adds it to the query keywords in the left frame. We can change the weight of thesaurus and log data by specifying it in the parameter boxes.

## Acknowledgments

We thank Minoru Kanehisa, Susumi Goto, Harumi Murakami, and Hiroyuki Kawano for their helpful comments and supports. This work is partly supported by a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science," from the Ministry of Education, Science, Sports and Culture in Japan.

## References

[1] Balabanovic, M. and Shoham, Y., Fab: Content-based, collaborative recommendation, *Communications of the ACM*, 40(3):66–72, 1997.

[2] Kawano, H., Mondou: Web search engine with textual data mining, *Proc. IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, 402–405, 1997.