GNOME: a sequence data management tool to access homology, motif, and other data analysis servers

Toshiyuki Niyama†, Takeo Tokimori†, Atsushi Ogiwara††, Ikuo Uchiyama††
Kenta Nakai†††

†Senri International Information Institute
8-47, Kakuda-cho, Kita-ku, Osaka 530, Japan
††Institute for Chemical Research, Kyoto Univ.
Gokasho, Uji, Kyoto 611, Japan
†††National Institute for Basic Biology
38 Nishigakoka, Myodaiji, Okazaki 444, Japan

Abstract

GNOME is a sequence data management tool through which users can efficiently access e-mail servers for various molecular biological analyses on Internet including GenomeNet. It supports BLAST/Fasta servers for homology searches, PROSITE/MotifDig servers for motif searches, and bget/blfind servers for DB entry retrievals. One of its most eminent features is that it can not only send e-mails for queries but also receive and manage e-mails for replies. In addition, its interface is very user-friendly. Therefore, it should considerably enhance efficient and profound analyses of newly-determined sequence data in both individual biological researches and large-scale genome projects.

1 Introduction

Homology searching for biological sequence database is the most popular technique for biologists who want to analyze sequences of unknown function. In many cases, however, programs for this purpose run on a local computer and require a database stored in the same machine. This means that users must keep databases in their local machines. As databases grow larger and are updated more frequently, maintaining databases personally becomes more laborious. Recently, with the development of computer networks, some organizations have begun database search services through electric-mail (e-mail) systems. In these services, specific programs called 'mailservers' receive query e-mails from users and return their results by e-mails.

One typical example is National Center for Biotechnology Information (NCBI) in the US, where the homology search service using the BLAST[1] program and the keyword search service using the IRx system are supported through Internet. In Japan, Institute for Chemical Research (Kyoto Univ.) and Human Genome Center (Univ. Tokyo) have started similar services through GenomeNet[2] since 1991 and 1992, respectively. Besides the same services as NCBI, they also provide the homology searching with the FASTA[3] program and the motif searching with the PROSITE[4] and MotifDig[5] databases. All of these sites maintain the non-redundant sequence database, which is the union of various public databases such as GenBank[6], EMBL[7], PIR[8], and SWISS-PROT[9]. Since it is updated in a daily basis, users can access the latest data at any time.

Although these services are potentially very powerful, there is some inconvenience: users must prepare query mail texts with proper parameters by themselves and must deal with their result e-mails manually. These processes tend to limit the usefulness of these services for most biologists who are not so familiar with computers. Moreover, these complex processes may severely restrict the efficiency of large-scale sequencing efforts in genome projects.

We have developed a user-friendly tool 'GNOME' to overcome such problems. Since GNOME adopts a popular graphical user interface (GUI) system, users can prepare query e-mails very easily, which allows them to concentrate on their own biological interests.
What is more important, GNOME can smartly manage the query results. Users engaging with a large-scale sequencing project need to check homology of massive data and are likely to receive too many result emails to handle without any management tool. GNOME is especially useful for such cases. With GNOME, users can systematically store their search results in named folders through its automatic classification mechanism. Moreover, these results can be summarized in various forms so that users can get insights from them easily.

2 Mailservers

 GNOME supports three kinds of mailservers: those for homology, motif and keyword searches. First, we give a brief overview of these servers. Figures 1(a)-3(a) show sample queries for each server. In homology and motif searches, an user sends an e-mail containing a query sequence and search parameters to the address of an appropriate server. Then the server runs a search program and sends back its result to the user by an e-mail as shown in figures 1(b)-2(b). In a keyword search, an user sends a query e-mail specifying keyword(s) and the database name, then its result will be received like figure 3(b). Query sequences for homology or motif searches are described in a so-called FASTA format in which a query must start with a name and comment line, marked with a letter '>', on its beginning, and sequence lines, coded in one letter, must follow. In homology searches, there are many kinds of parameters including program specification (i.e., fasta or blast), database specification (e.g., genbank), and search parameters such as a k-tuple for FASTA searches and a cutoff score for BLAST searches. In motif searches, the name of database, PROSITE or MotifDic, must be specified. In keyword searches, there are two statements: 'find' and 'get'. Find statement requires additional specifications for a database and keywords. Its result is the list of entry names that match these keywords. Get statement requires specifications for a database and entry names. Then the contents of specified entries are retrieved. The result of homology or keyword search servers is written in a flat-text style. The former includes a score histogram, a list of high-score entries, and sequence alignments.

The result of motif searches is also readable, but its graphics part is described in the PostScript language and requires some post-processing for visualization.

The graphics part consists of a schematic diagram of motif locations both in sequences and 3D-structures of proteins, if available.

3 Architecture of GNOME

 GNOME is a user-friendly tool for composing queries for above-mentioned mailservers and for processing their results. Figure 4 shows the configuration of its components. It can be roughly divided into two parts: 'search' and 'receive'. Search menu assists to compose queries for each server with proper parameters. There are three submenus corresponding to the three servers. Receive menu assists to get and process the results. It also has three submenus and in the homology and motif menus, there are additional submenus for each part of the result.

From the viewpoint of its management and operation, GNOME is also divided into two modules: the central data management module and the user front-end module. To establish multi-platform portability including personal computers, we adopted a server/client data communication model for the connection of these modules (Figure 5). The central data management module, also called 'GNOME DB Manager', works as follows:

1) Receives a query request from a client and sends it to an appropriate mailserver.
2) Accepts the results from that server and manages them for each user.
3) Manages folders to arrange the results for each user. All data are stored and managed in the DB Manager module and the role of the client module is only to display the data stored in the DB Manager. Such an architecture enhances the portability of the client module on various platforms, and it also guarantees the data consistency because the same DB manager always works wherever the client module exists.

4 Functions of GNOME

As already described, GNOME supports three kinds of mail servers. It also provides a simple sequence editor for sequence inputs. Each function can be called from its top level menu as shown in Figure 4.

4.1 Homology searching

For homology searches, full services supported by the mailservers of either NCBI or GenomeNet can
be utilized in GNOME. In the GenomeNet server, the FASTA program is also available as well as the BLAST program. Users can send requests and get their results in a same manner for each program and need not worry about their differences. Sending a request is easy using the search panel (Figure 6):
1) Choose Search⇒Homology from the top menu. The file selection panel appears.
2) Select a sequence data file to be sent. The search panel (Figure 6) appears.
3) Select a database and a program on the search panel. Change also the parameters, if necessary.
4) Push the Send button.
No description other than the query sequence itself is needed in the sequence data file since GNOME converts it automatically to the required format for each mailserver.

GNOME also provides users powerful functions to process search results. Figure 7 shows the various steps to examine the results. First, by choosing Result⇒Homology from the top menu, the list of search results is displayed (Figure 7(b)). After selecting any result and clicking the Detail button, the search result window is displayed (Figure 7(c)). In this window, entries found by the homology search are displayed in the order of their homology scores ('score-view'). Users can select two other types of display: the 'alignment-view' (Figure 7(d)) and the 'plain-view' (Figure 7(e)). In the alignment-view, similar regions detected in each entry are displayed in thick lines aligned with the corresponding regions of the query sequence. In this view, users can easily perceive the location and the length of each detected similarity. In the plain-view, the result returned from a mailserver is displayed directly. It can be used to get additional information such as the messages from the mailservers and the search time.

More precise information can be obtained from the score-view. Alignments between the selected entry and the query sequence are displayed by clicking the alignment button in the score-view window (Figure 7(f)). In this stage, users can directly retrieve the DB entry of interest by clicking the dbget button in the score-view window (Figure 7(g)), if the GNOME DB Manager is linked to the mailserver with TCP/IP.

4.2 Motif Searching

GNOME supports the motif search server on GenomeNet, which searches the query sequence for any motifs in the PROSITE or MotifDBIC database. Sending a request is performed in a similar manner with homology searching.
1) Choose Search⇒Motif from the top menu.
2) Select a sequence data file to be sent on the file selection panel.
3) Select a database and a program on the search panel (Figure 8). Change also the parameters, if necessary.
4) Push the Send button on the search panel.
The list of the results is displayed by choosing Result⇒Motif from the top menu (Figure 9(b)). Each result includes following information.
1) The schematic diagram indicating the location of motifs found in the query sequence. (Figure 9(d))
2) The schematic diagrams indicating the locations of motifs in the sequence family having the same motif as the query. (Figure 9(e))
3) The stereo diagram of DB proteins indicating the location of motifs, if their structures are stored in Protein Data Bank. (Figure 9(f))

4.3 Keyword Searching

GNOME also supports the entry retrieval servers through which users can get the contents of entries by specifying an entry name or an accession number ('get'), or can obtain the list of entry names the contents of which involve given keywords ('find'). Sending a request is performed in the following way,
1) Choose Search⇒DBget from the top menu and the search panel appears. (Figure 10)
2) Select a command (i.e., get or find), a database, and keywords.
3) Click the Send button.

Figure 11 shows the displays for their results. By choosing Result⇒DBget from the top menu, the list of results is displayed (Figure 11(a)). The contents of entries can be seen by selecting any line (Figure 11(b)).

4.4 Management of Search Results

Using folders, GNOME can manage its search results efficiently. That is, the result of each server can be classified and stored in a systematic way. Figure 12 shows its example for homology searching. Here, a result is about to be moved from the 'system/active' folder to the 'human/liver' folder. Not only moving, copying and deleting results, creating, moving, copy-
ing, and deleting folders are also allowed. These functions help users to manage their results easily.

5 Discussion

There already exist some tools that assist users to send queries by e-mail: Don Gilbert's 'GenBank Search' and Thon de Boer's 'mailfasta'. However, these tools do not have functions to deal with search results and users must use some general mail tool for this purpose. With GNOME, users not only can compose and send a query, but also can capture results, process and display them, and manage them with a consistent and easy-to-use GUI. These functions will strongly enhance the efficiency of researches that require large-scale data processing like genome projects. The area that GNOME covers, homology, motif, and keyword searching, are the top-most important services for many experimental researchers. In addition, using GNOME does not require any detailed knowledge on e-mail systems nor UNIX commands. Therefore, GNOME will also be very useful for general users who are not so familiar with computers.

 GNOME will be distributed to academic research organizations as a free software.

6 Acknowledgments

This project was organized by the Database Working Group in Human Genome Center (HGC), University of Tokyo and was supported by HGC. We thank Drs. Kaneshisa and Takagi of HGC for various suggestions and discussions. Dr. Akiyama of Institute for Chemical Research, Kyoto University kindly advised us on the 'Vibrant' library.

References


Figure 4. The component configuration of GNOME

Figure 5. The data communication model for GNOME
Figure 8. Sending a query for motif searching

Figure 9. Managing results of motif searching
Figure 10. Sending a query for DB-entry searching

Figure 11. Managing result of DB-entry searching

Figure 12. Example of folder operation