A Real-Time and Dynamic Biological Information Retrieval and Analysis System (BIRAS)

Qi Zhou[1, 2], Hong Zhang[1], Meiying Geng[2], and Chenggang Zhang[1] *
[1] *Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing 100850, China;*
[2] *The College of Applied Science, Beijing Polytechnic University, Beijing 100022, China*

The aim of this study is to design a biological information retrieval and analysis system (BIRAS) based on the Internet. Using the specific network protocol, BIRAS system could send and receive information from the Entrez search and retrieval system maintained by National Center for Biotechnology Information (NCBI) in USA. The literatures, nucleotide sequence, protein sequences, and other resources according to the user-defined term could then be retrieved and sent to the user by pop up message or by E-mail informing automatically using BIRAS system. All the information retrieving and analyzing processes are done in real-time. As a robust system for intelligently and dynamically retrieving and analyzing on the user-defined information, it is believed that BIRAS would be extensively used to retrieve specific information from large amount of biological databases in now days. The program is available on request from the corresponding author.

Key words: real-time information retrieval, sequence analysis, Entrez, BIRAS

With successively carrying out the human genome project and many other genome projects, the number of biological data and information sharply increased (*1*). To know the latest literatures and to find out the latest related homologous sequences have been very important in the field of bioinformatics. Researchers always encounter some unforeseen problems in their research. For instance, because a researcher does not retrieve the related literatures from Medline database in time, he might find some literatures of others already published and promulgated in the Medline, GenBank and other databases. In the field of novel gene cloning and function studies, researchers perhaps will face the same embarrassing situation that they do not know whether their concerned gene sequences or protein sequences have been applied for a patent or promulgated by others. Therefore, it is necessary to design an intelligent and dynamic searching engine for biological information retrieving to inform researchers. At this aspect, the National Center for Biotechnology Information (NCBI) has provided many related resources of bioinformatics and Entrez (http://www.ncbi.nlm.nih.gov/Database/index.html) (*2*), which is a famous search and retrieval system from databases at the NCBI. It greatly promotes the using of biological information. The Blast program is one of the widely used sequence homology analysis tools based on these databases (*3*, *4*). Users usually choose to connect directly to the Entrez Web Server in NCBI to retrieve literatures and gene sequences. However, for various reasons, users seldom connect to NCBI Entrez Web Server to retrieve data every day, so it sometimes causes unconsciously miss of some important information. To avoid such instances, this study will focus on the web-programming technique to design a Biological Information Retrieval and Analysis System (BIRAS). According to the user-defined retrieval requirements, BIRAS could automatically retrieve the latest biological information in the related database maintained by NCBI, and inform the users with the newest information by E-mail in each day. Thereby, BIRAS could furthest provide users with the services of real-time retrieval and analysis.

* **Corresponding author.**
E-mail: zhangcg@nic.bmi.ac.cn

The Scheme of Literatures, and Nucleotide Sequences and Protein Sequences Retrieval Based on Entrez

The aim of designing BIRAS is to make full use of the retrieval function of Entrez and timely retrieve the latest information of literatures and gene sequences. Usually, when we retrieve a keyword or term by Entrez Web Server in NCBI, we will first get a summary

list of literatures and gene sequences. If we find the necessary literatures and sequences, we will click their links and read the detailed information. Similarly, the retrieval process of BIRAS includes two steps. The first step is to send the keyword or term to the Entrez Web Server and gets the Identify (ID) list of the newly increased literatures and sequences. Secondly, it retrieves each ID and saves the retrieval information to local disk. Thus, BIRAS effectively retrieved user-defined updating information of the literatures and gene sequences.

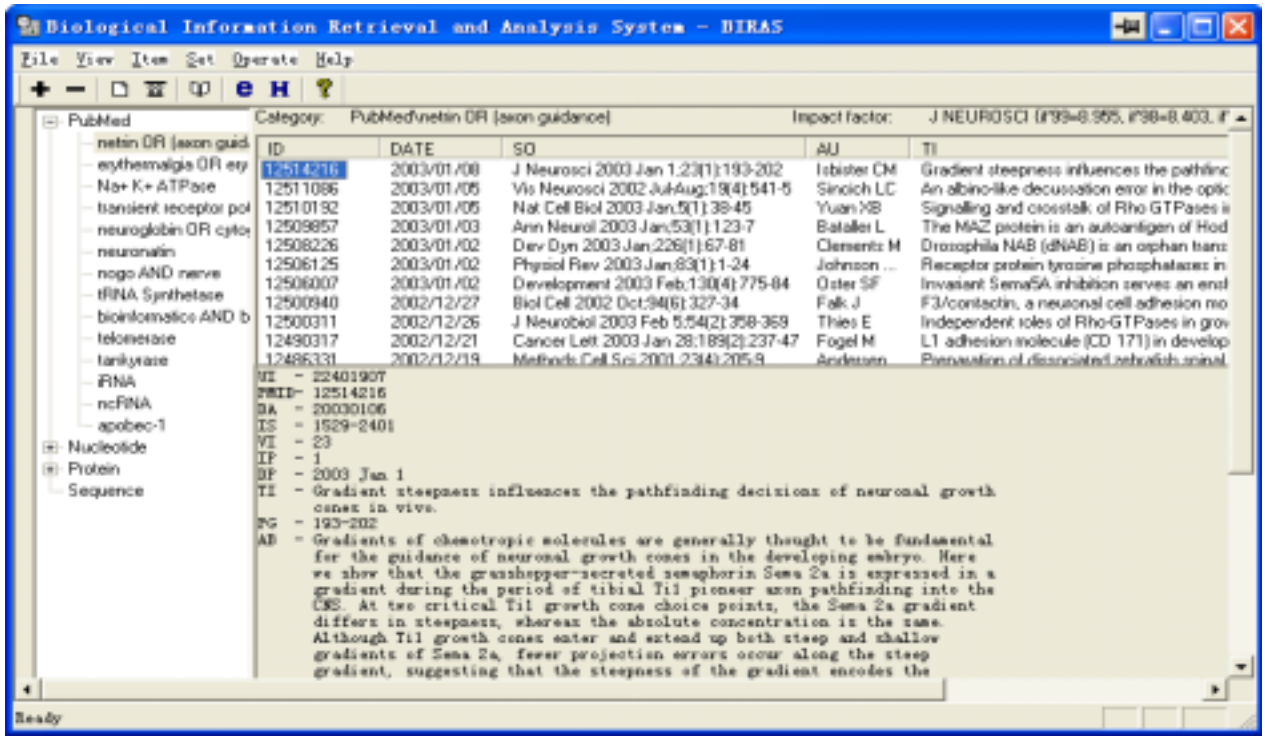

Fig. 1. Interface of BIRAS system. The left column indicates four kinds of items for user: PubMed (for Medline references), Nucleotide, Protein, and Sequence (for Blast analysis). The right top column indicates a short description of each retrieved result. For Medline reference, the impact factor was also displayed. The right bottom column indicates the detailed information of each retrieved result. Clicking on the top button "e" will open the retrieved file in HTML format in default web-browser. Clicking on the top button "H" could hide the program icon from the status bar into the system tray area.

The Medline literatures, nucleotide sequences and protein sequences retrieval section of BIRAS provides the complex term retrieval function of Entrez (http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html). Researchers can retrieve not only the simple single term but also the logical field composed of several terms with Boolean operators "AND, OR, NOT". Its rule is strictly coincident with the retrieval rule of Entrez. Furthermore, BIRAS also provides some common download styles (**Literatures**: Abstract, ASN.1, Citation, Medline, Summary; **Nucleotide Sequences**: ASN.1, FASTA, GenBank, Report, Summary; **Protein Sequences**: ASN.1, FASTA, GenPept, Report, Summary). Thereby, BIRAS provides a quite convenient retrieval mode for the researchers who concern with the special literatures and sequences. Besides, BIRAS provides the simple way for browsing the updating literatures and sequences, and shows impact factor of each journal as an especial item to make provide various kinds of information for the user (Fig. 1).

Scheme of Sequence Analysis Based on Blast

To the researchers who are studying structure and function of novel genes, it is very important for them to know whether there is any new sequences with high homology to their-concerned genes in the public GenBank database. Such sequences often come from homology genes of different species or alternative splicing forms of the same gene. Therefore, BIRAS provides an information retrieval mode based on the user-defined sequence. In other words, BIRAS automatically uses Blast program from the Web Server at NCBI to find out related gene sequences or protein sequences that are homologous to the sequence in GenBank. It is quite convenient for users to retrieve timely based on the homology of sequences. As far as the process of analysis, BIRAS adopts the QBlast's API to access the NCBI QBlast system (http://www.ncbi.nlm.nih.gov/BLAST/Doc/urlapi.html), and uses the corresponding Blast program of the NCBI QBlast system for analysis. When the NCBI QBlast process is finished, BIRAS will automatically save the result to local disk. After then, BIRAS does some necessary treatments with the analysis result and inform users of the latest increased homolog sequences in the GenBank database.

To automatically analyze sequence homology is an important function in the sequence analysis section of BIRAS. BIRAS allows users to set four common analysis parameters: Database, Expect, Descriptions, and Alignments (*4*, *5*). Researchers can set these parameters according to their demands in order to get the valid analysis result.

Program Design of BIRAS and the Interface

According to the above requirements, our designed BIRAS has the following interface (Fig. 1). In the Windows system, BIRAS can automatically run when Windows startup. Users can adjust the parameters of update conditions in the menu "Set/Update set", specify the server for sending E-mail in the menu "Set/Mail set", and define the same or different E-mail addresses for any retrieval item in the menu "Set/Mail Address". Thus, when the certain item retrieval finds out the updated information, BIRAS can send update information as attachments in HTML format to the user-defined E-mail addresses, in order to make users receive the latest information as early as possible.

Discussion

The outstanding characteristic of BIRAS program is that it can provide users with required information timely and dynamically according to defined requirements. Users can get the valid information easily enough, but need not download the whole Medline database or the whole nucleotide sequences and protein sequences database. Consequently, it obviously conduces to promoting the efficiency of their studies. At present, BIRAS merely concerns with the following contents: 1) Medline literatures retrieval. Namely, BIRAS timely informs users of the latest related literatures according to the user-defined terms, and shows the impact factor of each journal. 2) Nucleotide sequences and protein sequences retrieval, which means BIRAS timely retrieves the concerned nucleotide sequences and protein sequences according to the user-defined terms. If it finds out the new sequences, it will inform users of the information immediately. 3) Function analysis of special sequences. For this purpose, BIRAS will use the latest database to do some analysis such as homology analysis, structure and function analysis with user-defined sequence. The analysis result was also sent to users by BIRAS timely.

To a certain extent, information resources maintained by NCBI may be regarded as a dynamic data source. If user does not initiatively retrieve the Entrez Web Server at NCBI, the latest information will not be known by the users in time. Therefore, our reported BIRAS here is a quite important dynamic information retrieval and analysis system. In other words, BIRAS seems to be a worker who works quietly and always transfers user-concerned latest information timely. Therefore, BIRAS could provide users with the maximum information retrieval services in the future.

Compared with other similar systems, the real-time feature of literatures and sequences retrieval ability of BIRAS is quite outstanding. The Medline Workbench (http://www.medbench.org/; medbench updates Ver.3.1) allows users to order the information retrieval free service with the user-defined terms. However, in the practical test that respectively retrieves term "apoptosis" with BIRAS and Medline Workbench in the same time, BIRAS could find out 20 daily updated literatures in July 10, 2002. From

then on BIRAS has begun continuously retrieving the updating literatures every day. However, the Medline Workbench only sent a confirm E-mail to us within the first 24 hours, and did not contain any updating information. According to our experiences, in the daily using, the Medline Workbench merely sent about 1/4 number of literatures information at intervals. Therefore, its information real-time retrieval function is open to question. In fact, BIRAS directly connects to the Entrez Web Server of NCBI, and automatically retrieves updating information. Therefore, its performance is superior to those of some existing literature retrieval systems.

For the real-time literature retrieval, Mailing List is also a widely used means. For instance, *Nature, Science, J. Biol. Chem.*, and some other journals provide this kind of service. However, the Mailing List service has no definite keywords filtering feature. In other words, the Mailing List service of these magazines could only provide users the summary of the latest publications, but cannot filter the information automatically according to user-defined terms. When the user reads the whole content of the journals, it will take a lot of time to judge what they really need or not. Compared with this kind of time wasting, BIRAS can completely retrieve and effectively gather the users' concerned information from the databases of the NCBI according to the user-defined term. Obviously, BIRAS has the more practicability.

In terms of new sequences and their functions retrieval based on the sequence homology analysis, there are some related bioinformatic resources in Internet. For instance, the Swiss Shop server (http://www.expasy.ch/swiss-shop/) allows user to specify the protein sequence and do homology analysis, and then returns the analysis result to user by E-mail. However, the system cannot deal with the nucleotide sequences. The EMBL database has ever provided the similar service, and then it stopped the service for some reasons. Compared with these resources, our sequence homology analysis function of BIRAS includes both the nucleotide sequences homology analysis and protein sequences homology analysis. Furthermore, it can carry out analysis using all databases in NCBI including Sequences Databases derived from the Patent division of GenBank, Database of expressed sequence tags (EST), and Databases of High Throughput Genomic Sequences (HTGS) and so on. Users can freely use these databases by setting different parameters.

For realizing the characteristic of running in various operating systems (Windows/ Linux/Macintosh *etc.*), the core code of BIRAS is written in standard C language, and then joins the corresponding management software based on the individual characteristic of the various operating systems. At present, the Windows version of BIRAS can run in Windows 9X/Windows NT/Windows 2000/Windows XP normally. Of course, the computers installed with BIRAS should have connected with the Internet directly.

In conclusion, through many strict tests, BIRAS can retrieve the information of user-concerned literatures and sequences timely and effectively from the Entrez Server at NCBI. In fact, if you can retrieve and gain the information by using the Entrez, BIRAS can automatically retrieve them and inform users in time. Thereby, it is believed that BIRAS would be used to retrieve specific information from large amount of biological data in now days.

References

1. Cotton, R.G. 2002. The Human Genome Project and genome variation. *Internal. Med. J.* 32: 285-288.
2. Jenuth, J.P. 2000. The NCBI. Publicly available tools and resources on the Web. *Methods Mol. Biol.* 132: 301-312.
3. Pertsemlidis, A. and Fondon, J.W. 3rd. 2001. Having a BLAST with bioinformatics (and avoiding BLAST phemy). *Genome Bio.* 2: REVIEWS2002.
4. Altschul, S. F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
5. Zhang, C.G. and He, F.C. 2002. *Bioinformatics: method and practice,* pp. 249-257. Science Press, Beijing, China.

This work was partially supported by Chinese High-tech Program "863"(2002AA234021), National Science Fund for Distinguished Young Scholars (30128010), Chinese National Natural Science Foundation General Program (30200154, 30100049, 39900041 and 39900074), General Program of Natural Science Foundation of Beijing, China (7002030) and Initiative Foundation for Scientific and Technological Innovation of Academic Military Medical Science (0102001, 9905105).

Received: 19 November, 2002
Accepted: 9 January, 2003