1998 International Computer Symposium
Workshop on Software Engineering and Database Systems
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

Web-Based DNA Sequence Annotation Tools for Functional Genomics

Chung-Shyan Liu
Dept of Infor & Computer Engr
Chung Yuan C. University
Chung-Li, Taiwan
liucsmars.ice.cycu.edu.tw

Wei Kuo
Dept of Infor & Computer Engr
Chung Yuan C. University
Chung-Li, Taiwan
weikuo@venus.ice.cycu.edu.tw

Lan-Yang Ch'ang
Institute of Biomedical Sciences
Academia Sinica
Taipei, Taiwan
lychangibms.sinica.edu.tw

Wen-Chang Lin
Institute of Biomedical Sciences
Academia Sinica
Taipei, Taiwan
wenlinmail.ibms.sinica.edu.tw

Shih-Ping Tung
Dept of Mathematics
Chung Yuan C. University
Chung Li, Taiwan
sptung@cchp01.cc.cycu.edu.tw

Abstract

In this paper, we will present the design and implement of two web-based tools for DNA sequence data annotation. Our tools integrate functional genomics sequence data from different sources and in different formats. One tool handles the e-mails that contain DNA sequence data from sequencing machine and lets the user view and edit the sequences. The sequences may then be stored in a local data warehouse or be sent to a remote database to search for matches. Another tool handles the database search results and lets the user annotate the matched sequences before storing the data into a database, called CR Base, for further exploration. Our tools use the Internet Web Browser as the only user interface, so that the user may use the tools to access CR Base and to perform data manipulation on Internet.

1 Introduction

Information technology has always played an important role in the development of genetics and molecular biology. Information technologies helped to develop and provided many useful algorithms, working programs, and software tools, like GCG and BLAST, for molecular biologists to deal with sequence analysis problems in computational biology, such as mapping, alignment, DDP, and so on [1][3]. Information technologies also helped the development of some very large sequence databases, such as GenBank, TIGR, and SwissProt, and provided software tools for data management, data analysis, and database search [5][11][12]. With the increase of throughput of sequence data generation, in part due to Human Genome Project, it is evident that the interpretation of the sequence data to obtain the biological signals, the genes, and other sequence characteristics will be the central issue of bio-informatics research. The development of sequence analysis algorithms will not play a major role in the future of bio-informatics research. The next major step in molecular biology and genetics is not the generation of sequences, but the analysis and interpretation of data [7][8][12].

The analysis and interpretation of a new segment of DNA sequence includes identifying and characterizing the sequence. Such work involves comparing the DNA sequence with known protein sequences and other DNA sequences that are stored in various databases. The databases are distributed and managed by different institutes and the sequence data stored in different databases are usually in different formats. The new DNA sequence, which is determined by a sequencer, is never perfect and needs manual verification edit before sending out for database search. Also, the results from database search must be manually verified and edited before being stored in a local database. It is obvious that under current practice, a molecular biologist will spend a substantial part of time in such routine, but vital work.

As Internet is gaining popularity, more databases and labs are putting the sequence data on the Internet. With the capabilities of graphical user interface, web-browser is now the dominating user interface to the Internet. Thus, it is expected that the annotation software should also provide access to the Internet and preferably use web-browser as the user interface. However, an annotation tool must provide more functionality than those needed for surfing the Internet. For example, an annotation tool must have the ca-

pabilities to query and search databases, to edit and annotate sequence data, or to perform other functions. Such functionality will be provided using CGI or Java applets.

In this paper, we will present the design and implement of two web-based tools for DNA sequence data annotation. Our tools integrate functional genomics sequence data from different sources and in different formats. One tool handles the e-mails that contain DNA sequence data from sequencing machine and lets the user view and edit the sequences. The sequences may then be stored in a local data warehouse or be sent to a remote database to search for matches. Another tool handles the database search results and lets the user annotate the matched sequences before storing the data into a database, called CR Base, for further exploration. Our tools use the Internet Web Browser as the only user interface, so that the user may use the tools to access CR Base and to perform data manipulation on Internet.

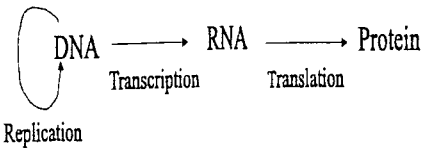

Figure 1: The information flow of central dogma in biology

2 Background

The heredity signals of living beings are carried in and determined by the DNA sequences, which are composed of four different nucleotides, Adenine(A), Guanine(G), Thymine(T) and Cytosine(C). The DNA sequences are in double helix structure. The DNA sequences are replicated and then transcripted into RNA secondary strucutres. The RNA structures are then translated into amino acid that constitute protein [3][4]. This process is referred to as *central dogma* and is shown in Figure 1. However, except in *prokaryotes*, that are oganisms without a nucleus, not the whole DNA sequence is used for transcription. The part of DNA sequence used for transcription is called coding region and that part constitutes the genes of a living being. During the development period of a living being or when being acquired genetic diseases, the disease-genes (the coding region) expressed will be different from that of normal one. Therefore, the diseased tissue generates unwanted protein or fails to generate the necessary protein, both are characterized by the symptons of the disease.

Functional genomics deals with the expressed genes and the functionality of that sequences. It consists of two areas: mRNA transcription and protein structure. In this paper, we deal primarily with transcription. Information technologies will play an even more important and essential role in functional genomics, from gene annotation to gene expression to the establishment of expressed gene anatomy. For example, computer tools is necessary for profiling gene expression patterns, as a single cell type has 15000 to 20000 expressed genes, while there are around 100,000 genes in human. Without a software tool, it is almost impossible to have any good analysis results.

To obtain useful information from sequence data, one has to annotate the sequence data, in order to

1. find putative biological signals and understand sequence characteristics,
2. discover putative genes, for instance, the target genes involved in the disease process, which can be used for diagnosis and for finding lead compounds or drugs, and
3. maintain the information of gene profiles and combinatorial chemistry for drug discovery.

The sequence data to be annotated may come from different sources:

1. From automatic sequencer. In this case, the sequence data has to be checked for quality.
2. From database search, such as GenBank or TIGR. In such case, one needs to check the percentage of match and the consensus of sequence data.
3. From local laboratory, such as EST (Expressed Sequence Tag) or partial cDNA sequences.

The annotation may be performed in different ways:

1. Manual edit,
2. Automatic edit, if the edit patterns are fixed, or
3. Using data mining to find out some useful patterns.

But annotation is not an easy task. Although manual edit is still the most commonly used method, with the current throughput of sequence data generation and the constantly updates required of the annotation, due to database update or new experiment results, it is evident that manual edit is becoming an infeasible method. Thus, how to provide automatic tools to support annotation of sequence data has become one of the most important issues in Bio-Informatics research.

An automatic tools for annotation must deal with the following problems:

- Sequence data may come from different sources, as mentioned previously.
- As sequence data come from different sources, they are in different format.
- The sequence data may come in different ways, such as e-mail, ftp, local files, or local database queries.

Also, it is expected that the annotation tools should have a friendly user interface and should be very flexible, so that it can be easily modified to satisfy requirements in the future.

3 Workflow Analysis

As indicated in previous section, a segment of DNA sequence data needs to be annotated so that useful biological signals and information may be obtained. Since each laboratory has its own set of sequence data, which may be obtained from experiments or from previous database search, and each laboratory has its own annotations of the data sequences, it is desirable that each lab maintains a local database, which is called data warehouse in this paper. Also, each lab may have obtained some DNA or protein sequences that their functions and characteristics are well known and experimentally verified, and the lab wishes to share this information with others or to publicize the information. In such case, the sequence data are put to a public database, called CR Base in this paper, that others may access or search on the Internet. The overall roadmap of our lab is shown in Figure 2. But in this paper, we will only consider the part that is in light color. Other part of the roadmap is out the scope of this paper.

The overall processing steps is as follows:

1. A sample of DNA segment, which is resulted from experiments in the lab, is sent to sequencing machine, called *sequencer*.
2. The sequencing machine sequences the sample and sends the DNA sequence data back by e-mail. This step may take several days.
3. The DNA sequence is manually verified and then compared with the sequence data of a local database, called CR base in this paper. If there is a good match, the DNA sequence is annotated and then stored into the CR base.

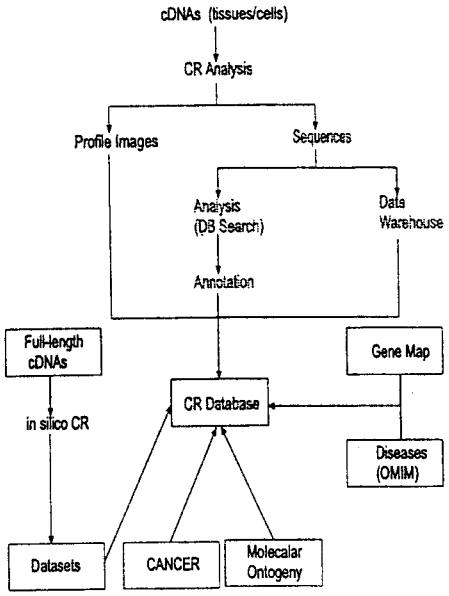

Figure 2: The roadmap of functional genomics.

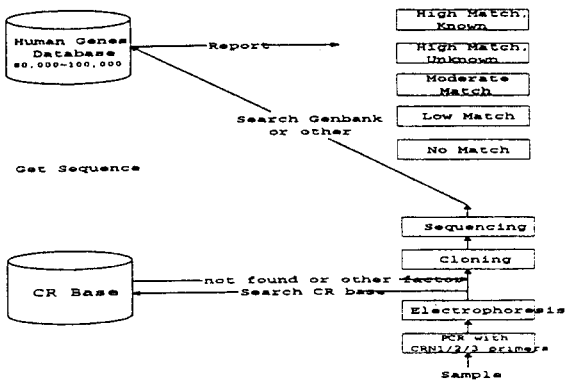

Figure 3: The workflow and processing step of sequence data from a sequencer.

4. If there is no good match or further characterization of the DNA sequence is required, the DNA sequence is sent to other remote databases, such as GenBank or TIGR, to search for matches. Before sending the DAN sequence for database search, the sequence data must be converted into the required format first.

5. The results of database search will be sent back in e-mails or in html if the request is done by http. This step may range from several minutes to several days. If there is no response within some time, say three days, a re-submission of search will be needed.

6. The results of database search is manually checked and annotated and then inserted into local CR Base. Since the results from database search will be in different format when searching different databases, it is necessary that the results be converted before being stored into CR Base.

The overall processing steps are shown in Figure 3.

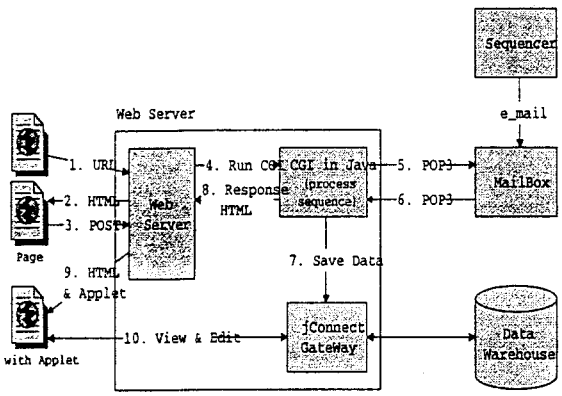

Figure 4: The workflow of processing DNA sequence data from a sequencer.

From the analysis, we found that the steps of processing sequences from sequencer and processing database search results are the most elaborate and laborious steps. These steps also need continuous manual attention, since the completion time is highly unpredictable. Thus, it is hoped that these steps can be automated as much as possible. Currently, although we have developed some window-based programs to support the routine work as outlined above, those programs must be invoked separately. Also, since the temporal duration of the workflows are quite long (may take several days) and the time for some tasks is non-deterministic(such as sequencing or database search), this solution requires the users to check the compeltion of each program and then to handle the data convertion between different tools manually. Thus, a better integrated solution is needed.

To process the DNA sequence data from the sequencer, the user must read the e-mails from the sequencer and then verify, or edit if necessary, the DNA sequence data. Then the user will store the sequence in data warehouse or submit the sequence to database to search for possible matches. To store a sequence data in data warehouse, the user selects certain fields from the sequence data and then invokes another program to store the data. To submit the sequence data to a remote database for searching a match, the user must convert the sequence data into the format required by the database. In Figure 4, the workflow of our tool that processes the e-mails returned by the sequencer, allows the user to view the results and then to annotate the sequence data, and then lets the user store the sequences and annotations into data warehouse, is shown. The processing steps are marked numerically.

There are two ways of submitting a sequence to a database, such as GenBank [10] in our case, to search for possible matches: using e-mails or using web-pages. Since the results sent back by GenBank do not contain any identifying information and may not be in the same order of submissions, if the requests are done by e-mails, it will be difficult to process the results automatically if more than one requests are submitted at the same time. Thus, only one request will be submitted each time, using http, and next request will be submitted only if the response of current request has been received. Also, as the reply time is unpredictable, it is necessary that the user be notified when the responses come in. The results returned from database search contain information about other sequences that match the submitted sequence with scores. The matched sequences are classified into *high* match, *low* match, and *no* match groups. Two sequences are likely to have similar functionality if they are *high* matched. As the database search results must be verified and annotated by the user, such processing can not be fully automated. The workflow for processing the results from searching or mining databases is shown in Figure 5, where the processing steps are marked numerically.

4 Web-Based Tools

In this section, we will present the implementation of two web-based tools for processing the e-mails from

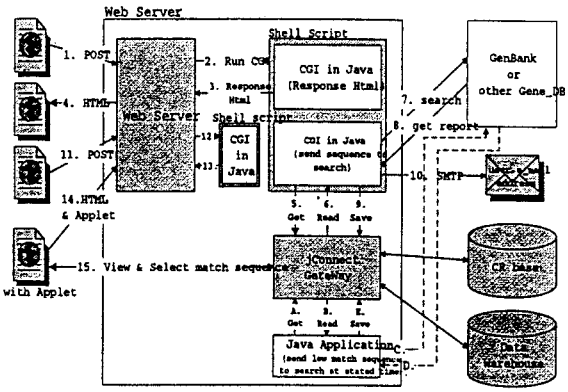

Figure 5: The workflow of processing database search results.

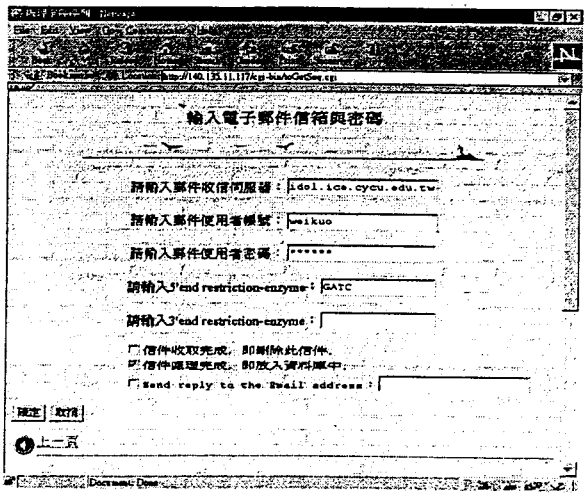

Figure 6: The initial screen of e-mail processing.

sequencer and the results from database search. The only user interface of our tools is a web browser. Although the processing and computations may need to be performed separately, the user interface is integrated.

The tool that processes the e-mails from sequencer has two main pages on the browser. The first page, shown in Figure 6, processes the e-mails sent back by the sequencer and the second page, shown in Figure 7, deals with the annotations. Currently, the web-server and mail-server locate on the same machine, but it is possible that they may locate on different machines. Since the client machine, where the web browser sits, is usually a PC, the mails need to be retrieved from the mail-server using pop3 protocol [2]. After a mail is retrieved from the server, the user may view and edit the sequence data and then store the data into the data warehouse. Under our current inplementation, the client site handles user interface only, and most of the processing and computations are performed at the server site. We used jConnect, as shown in Figure 4, as Java to database interface, which is Sybase in our case.

In Figure 8, the screen of the tool that enables database search and processes the database search results is shown. As stated before, the sequence data are submitted to a remote database for searching matches using http protocol, after the required data conversion. Since the reponse time from database search is uncertain and may be lost, a timer is set and when timeout, a re-submission is performed. Also, the responses from some databases, such as GenBank, do not contain identifying information, the next request can not be sent before the previous response is returned. Our tool is able to handle a batch of requests and responses as a background thread and to alert the user of the responses. The user may use the tool shown in Figure 8 to annotate the sequences and then store the sequences in CRbase.

5 Discussion and Future Research

From the development of our web-based tools, some important points are observed.

First of all, one should not confuse a web-based application with home page desgin, although home page design is a crucial part of a web-based tool. The desgin of a web-based tool is much more complicated, it requires that the process model of software development be strictly followed. It is important to note that a web-based tool is essentially a distributed system, and the difference between a web-based tool and a traditional distributed application is the user interface. Thus, the design methods for traditional client/server model can

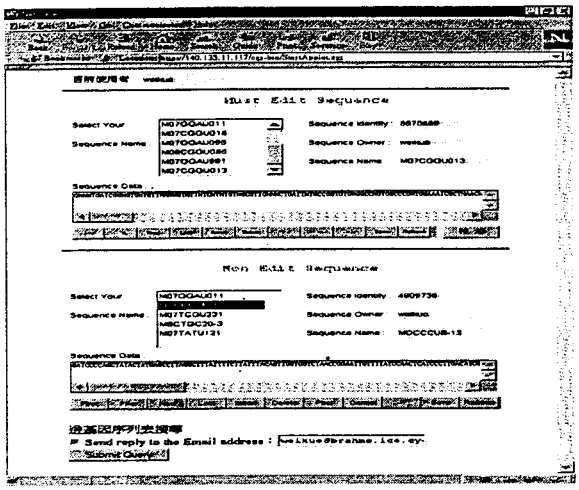

Figure 7: The processing of sequences from sequencer.

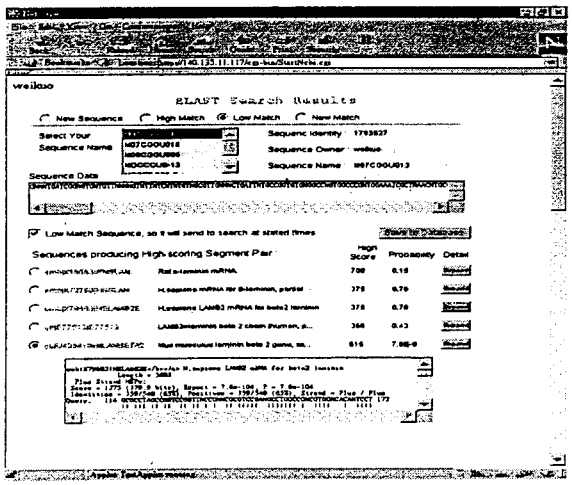

Figure 8: The processing of database serach results.

be used.

However, Current web-based applications enriches the traditional client/server paradigm. In traditional client/server model, a client makes requests and the server responds to fulfill the requests. The client is usually a simpler machine that only performs trivial tasks, such as display and user interactions, while the server is a more complicated machine that performs most of the processing and computations. But as PC and desktop workstation become more powerful, the client machine may perform more complex tasks and processing. Thus, one of the important design objective is to partition the processing between the clients and the servers so that their loads will be balanced. But in traditional client/server model, when an application is re-designed to achieve the optimal performance, the programs on the client machines need to be re-installed and this will be a maintenence burden.

A web-based tool has the same advantage of a window-based tool, as both have similar graphical user interface. A window-based application is normally event-driven, which is good for user interface since the user can invoke the functions as needed. But this is not necessarily good for a workflow application, since now here-there jumps are allowed and thus will be difficult to keep track the processing steps. Also, in a window-based tool, the information are scattered in different components that must be invokded separately. User must take explicit actions to integrate the information. In a web-based tool, the information may be integrated behind the scene at the server side and thus transparent to the users.

6 Conclusion

We have presented the design and implement of an automated software system that integrates the functional genomics sequence data, based on CR approach, from different sources and in different formats. Our automated annotation system can take sequence data either from sequencer or from global database search. Our system enables the user to establish CR Base, to annotate sequence data and to profile expressed genes. Our system uses the Internet Web Browser as the only user interface, so that the user may use the system to access CR Base and to perform data manipulation on Internet. The most important lesson that we learned from this work is that a web-based tool enrichs the traditional client/server computing paradigm while at the same time has the friendly graphic user interface. It is also expected that as more interaction tools are developed, it will be easier to develop a web-based application in the future.

References

[1] E.S. Landef and M.S. Waterman ed., "Calculating the secrets of Life: Applications of the Mathematical Sciences in Molecular Biology", National Acdemy Press, 1995

[2] J. Myers and M. Rose, "Post Office Protocol, Version 3", Nov 1994, RFC 1725.

[3] M.S. Waterman, "Introduction to Computational Biology, Maps, Sequences and Genome", Chapman & Hall, 1995.

[4] J.D. Watson, M. Gilman, J. Witkowski, and M.Zoller, "Recombinant DNA", 2nd Edition, Scientific American Books, 1992.

[5] *Science, Genome Issue*, Vol. 270, No.5235, Oct.20, 1995

[6] E. Shakhnovich and A. Gutin, "Implications of Thermodynamics of Protein Folding for Evolution of Primary Sequences", *Nature*, Vol-346, 1990, 773-775.

[7] C. K. Peng, et al, "Mosaic Organization of DNA Nucleotides", *Physical Review E*, Vol-49, No 2, 1994, 1685-1689.

[8] C. K. Peng, et al, "Long Range Correlations in Nucleotide Sequences", *Nature*, Vol 356, 1992, 168-170.

[9] R. Mantegna, et al, "Linguistic Features of Noncoding DNA Sequences", *Physical Review Letters*, Vol-73, No 23, 1994, 3169-3172.

[10] http://genome.nhgri.nih.gov

[11] Proceeding of Fourth International Conference on Bio-Informatics and Genome Research, San Francisco, CA, Jun 1995.

[12] Proceeding of Fifth International Conference on Bio-Informatics and Genome Research, Baltimore, MD, Jun 1996.