한국지능시스템학회 논문지 2008, Vol. 18, No. 1, pp. 27-31

다중서열수집 및 변환을 위한 효과적인 바이오인포메틱스 도구

An Effective Bioinformatics Tool for Multiple Sequence Acquisition and Translation

이혜리[1], 이승희[1], 이건명[1], 김성수[1], 이찬희[2], 이성덕[3]
Hye Ri Lee, Seung-Hee Lee, Keon Myung Lee, Sung-Soo Kim, Chan-Hee Lee, Sung-Duk Lee

[1]충북대학교 전기전자컴퓨터공학부
[2]충북대학교 미생물학과
[3]충북대학교 정보통계학과

요 약

많은 바이오인포매틱스 관련 데이터베이스와 도구가 네트워크를 통해서 제공되고 있고, 이들을 효과적으로 활용하면 생물학적 분석을 적은 비용으로 우수한 결과를 얻을 수 있다. 이 논문에서는 주어진 질의에 대해서 잠재적으로 관련된 DNA 서열 정보를 획득하고, 분석자가 관심있는 항목을 선택하면, 선택된 항목에 대한 모든 DNA 서열 정보를 확보하고, 이들에 대해서 아미노산 서열로 자동변환하여 ORF라는 정보를 활용하여 가장 가능성이 큰 것을 추천하는 도구를 소개한다. 해당 도구에는 웹 로봇 기법과 ORF 검색등을 위한 생물학적 지식을 활용한다.

Key words : bioinformatics, web robot, sequence analysis

1. Introduction

Various valuable databases and analysis tools are available for molecular biological studies and bioinformatics over the network free of charge for the analysts. A bioinformatic analysis usually requires to use multiple databases and/or analysis tools. Even though individual tools and databases are efficient, it is sometimes bothersome to use together multiple databases and tools. An effectively organized tool to make use of other tools and databases could greatly contribute to reduce analysis efforts.

With the help of high throughput sequencing technologies and efficient computational tools, the genomes for dozens of organisms have been successfully sequenced and archived into biological databases. Many researchers have been paying their attentions to post-genome studies such as identifying the functions of genes, modeling the interaction networks like gene regulatory networks, signal transduction networks and metabolic pathway networks, and so on.

Once a sequence is obtained from some biological treatment, the biological analyst tries to deduce its structural, functional and evolutionary relevance by evaluating its similarity and difference in DNA base-level or amino acid-level with respect to other sequences of interest.[1] The biological analysts usually collect the existing sequences and corresponding annotation information from public biological databases like NCBI[2] and carry out the designated analyses with web applications and stand-alone applications. Various bioinformatics tools have been being developed to find optimized or approximated solutions, to perform complicated computations, to carry out effectively the analysis tasks manually done by the analysts.

This study is concerned with the analysis situation in which an analyst needs to get multiple DNA sequences from a public database and to convert them into amino acid sequences. In order to accomplish this task, she first has to enter GenBank Accession Number or Geninfo Identifier into the NCBI GenBank database[3], extracts out the DNA sequence sections from the retrieved results and store them into a file. She now translates one by one the collected sequences into amino acid sequences manually or with the help of translation program like ExPASy's *Translate*[4], and then stores them in a file in a required format. This paper presents an intelligent tool which has been designed and implemented to perform the above-mentioned whole steps at a time.

This paper is organized as follows: Section 2 briefly reviews some related works on DNA sequence translation tools. Section

접수일자 : 2007 년 11 월 13 일
완료일자 : 2008 년 2 월 11 일
This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (The Regional Research Universities Program/Chungbuk BIT Research-Oriented University Consortium)

3 presents the proposed approach and the developed system in term of system architecture and its functionality. Section 4 discusses how effective to use the proposed tool in the biological analysis Section 5 draws the conclusions and future works.

2. Related Works

Various databases and analysis tools are available over the network in the bioinformatics domain. Representative databases widely used to get biological information include GenBank, UniProtKB/Swiss-Prot of EMBL[5] and PIR-International[6]. They are equipped with the query interfaces and provide retrieved results in a web page or in various standard format like ASN.1, XML and FASTA.

Each record in these databases consists of various fields including sequence, annotation, organism, reference, and so on. The databases have partly overlapped records for the same objects each other even though they have different focuses, e.g., some for DNA, others for protein. In the biological databases, each record has a unique identifier called access number or other name in the specific database, with which the records are linked across the databases.

GenBank is one of the most frequently used databases to which newly identified sequences are generally registered. It contains data about DNA and protein sequences, genome maps, molecular modeling database, and documentary comments[3]. For GenBank, NCBI provides a Web query interface program, called Entrez, one of which functions allows to retrieve DNA sequences along with its related information with GenBank Accession Number or Geninfo Identifier[2]. Entrez can publish the retrieved results in various formats like plain text, XML, and so on, from which users have to extract the DNA sequence.

ExPASy(Expert Protein Analysis System) provides a tool called *Translate* which translates a given DNA sequence into its corresponding amino acid sequence[4]. ExPASy is a system operated by SIB(Swiss Institute of Bioinformatics) to provide proteomics-related information. *Translate* allows to convert a DNA sequence at a time. Therefore, to handle multiple DNA sequences its users have to do bothersome cut-and-paste works as many times as the number of sequences to be translated.

When a user collects multiple DNA sequences of interest and translates them into the corresponding amino acid sequences, she carries out the following steps at this stage of available technology: She retrieves DNA sequences by their accession numbers with NCBI Entrez, manually extracts DNA sequence parts from the retrieved results and transforms into FASTA format, then feeds DNA sequences into ExPASy's *Translate* one by one while cut-and-pasting the translated sequences into a file. This series of steps is a time-consuming and bothersome task. From these observations, we have developed a system which takes care of the whole process with the minimized user's intervention. One of the purposes in the bioinformatics tool applications is to reduce the analysis cost by reducing the bothersome tedious tasks.

3. The Developed System

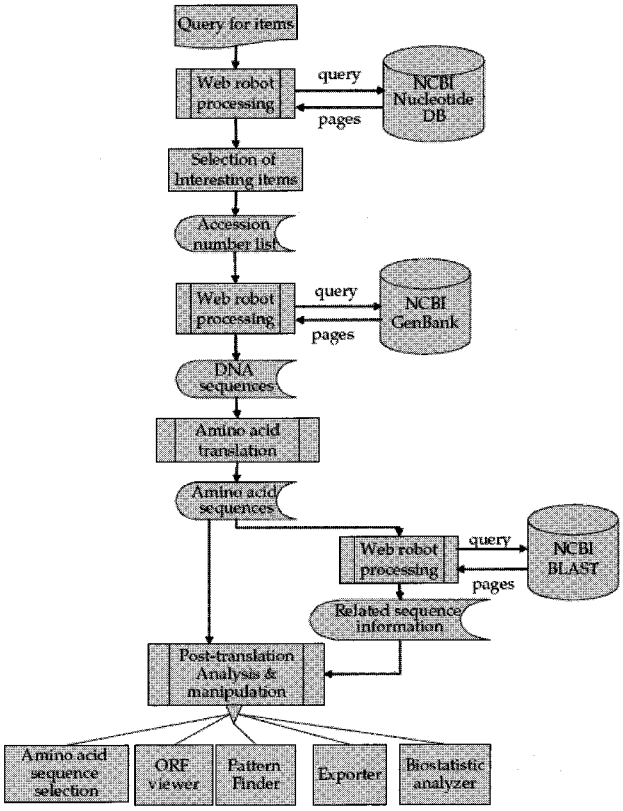

Figure 1. The Work Flow of the Developed System

For automatic multiple DNA sequence acquisition and translation, a system has been developed which consists of the DNA sequence acquisition module, the translation module, and the auxiliary tools such as Pattern Finder, Biostatistics Viewer, and Exporter. The DNA sequence acquisition module has been implemented using a web robot technique which plays role of collecting the DNA sequences from NCBI GenBank when a set of accession numbers for DNA sequences is given. The translation module is charge of translating the collected DNA sequences into amino acid sequences in which 6 amino acid sequences are generated for each DNA sequence according to the possible reading frame positions and directions, and then the most plausible one is recommended. Pattern Finder allows to search for a specific subsequence from the selected DNA sequence. Biostatistics Viewer shows some basic statistical information for the sequences. Exporter plays role to export the processed sequences into a file. When a user makes use of this

system, she can ask to collect the DNA sequences from the NCBI GenBank by listing out their accession numbers or can provide directly the DNA sequences to the translation module. The DNA accession numbers to be collected are listed in a file and the file is handed over the system as the input file. The files to store the DNA sequences collected by the user are edited in the FASTA format, and the system provides a functionality to handle these files to translate into amino acid sequences.

For each DNA sequence, the system takes into account 6 possible translations and recommends the most likely one at the first place and enables the user to confirm it or to choose other one from the translated sequences through the graphical user interface. The confirmed amino acid sequences are exported into a file in the FASTA or XML format. Figure 1 shows how the system works for the multiple DNA sequence translation.

The system has been implemented into a standone application for the Windows environment using Microsoft Visual Basic components. Due to its necessity to access the NCBI GenBank, the application should be deployed in the computers which are Internet-connected.

3.1 The Query Robot to Gather Information

The query robot plays role to receive the analyst's query, send it to NCBI Entrez, receive the reply and parse it to extract the information of interest. It shows the retrieved information related to the query keywords and allows the analyst to select the items of which DNA sequence is retrieved from the NCBI nucleotide database.

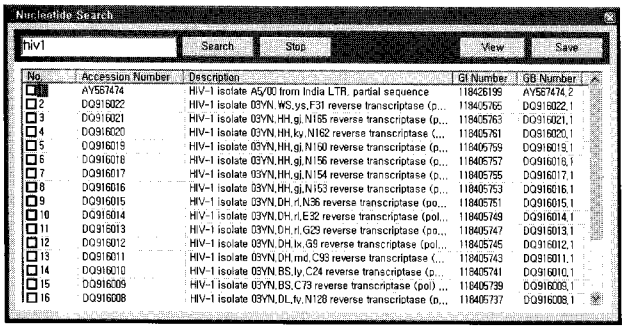

Figure 2. The query robot interface. The analyst enters keywords in the edit box and the robot displays the retrieved results on the lower window plan on which he can select the item by checking the checkbox on the very left in each line.

3.2 The Web Robot to Collect DNA Sequences

To collect multiple DNA sequences of interest from the NCBI GenBank, a web robot has been developed which interacts with the NCBI Entrez. The web robot first uploads the file containing the DNA accession numbers, then retrieves the DNA sequences corresponding to the accession numbers one by one from the GenBank, and the collected sequences are maintained in an internal data structure and displayed through the tree-view directory graphical component.

For each accession number, the web robot creates a query to be delivered to the Entrez's CGI program. In response to each query, the CGI program retrieves the corresponding record from the GenBank's Nucleotide database and sends back the retrieved result. Then, the robot parses through the received page and extracts out the DNA sequence. The collected DNA sequences are maintained in a FASTA format data structure and are used in the following translation work. The web robot helps the analysts avoid sequence-wise bothersome and time-consuming manual interactions with the Entrez system to collect the sequences. It could contribute largely to the analysis time reduction.

3.3 The DNA Sequence Translation

A DNA sequence is made of base characters each of which A is for adenine, T for thymine, G for guanine, and C for cytosine, and encodes genetic information. Each consecutive three bases, called a codon, could encode an amino acid[7]. There are 20 kinds of amino acids corresponding to 64 possible codons. Each codon has a single letter code for its amino acid, e.g., 'M' for methionine[8].

For a DNA sequence, there are 3 possible reading frames in each the directions from 5' to 3', and from 3' to 5'. A reading frame is the way in which nucleotides are read in groups of three to specify a code. The developed system generates 6 amino acid sequences corresponding to each reading frame. When a DNA sequence is selected, the graphical user interface shows its amino acid sequences and allows to choose one out of 6 sequences, as shown in Figure 3.

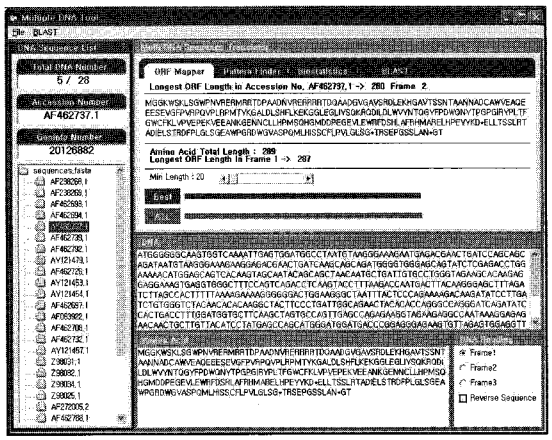

Figure 3. The Amino Acid Sequence Viewer

For each reading frame, there is a sequence of codons, called ORF(open reading frame), beginning with an initiation codon (i.e., ATG) and ending with a termination codon (e.g., TAA, TGA, TAG). An ORF is a potential coding area which encodes a gene and thus is used to compose a protein when it is expressed. In biology, an ORF plays important role in determining whether

its sequence encodes potentially a protein composition information by carrying out homology analysis at the amino acid level, or as a yardstick when evaluating a coded protein's size or molecular mass.[9] A meaningful ORF is usually the longest reading frame containing no in-frame stop codons.

In order to search for the longest ORFs, the developed system takes into account all possible combinations between M, which is the initiation codon, and *, which denotes a termination codon. With the threshold of the minimum ORF length which can be controlled by the analysts, the longest one of size greater than the threshold is selected as the ORF at the corresponding reading frame.

To tell the most likely amino acid sequence for a DNA sequence, the developed system determines the amino acid sequence with the largest ORF among the 6 possible translated amino acid sequences. The most likely sequence is displayed with which the corresponding check box is in the set state in the amino acid sequence viewer when an accession number is selected on the access number list window. This recommendation helps the analysts confirm the translated amino sequence with little effort, where they can select different translation other than the recommended one. The confirmed sequences are later exported into a file.

3.4 The Web Robot to Communicate with BLAST

The saved sequences could be later used as a query sequence for the sequence similarity-based retrieval services like BLAST. For the selected sequence, the developed tool's web robot access the BLAST tool over the network, it collects the highly matched sequences with the help of BLAST. The developed tool reorganizes the received data and provides them in a concise form for the analysis as shown in Figure 4.

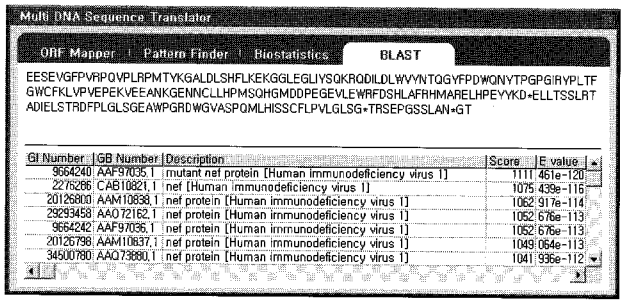

Figure 4. The BLAST Result Viewer

3.5 The Other Functionalities

The ORF Mapper is the functionality to display the recommended ORF for the selected amino acid sequence. In its window display, the possible ORFs' positions are depicted by the bar graphs. The threshold for the minimum length of ORFs can be controlled by the analysts with the graphical sliderbar interface. The window displays a simple statistics about the amino acid sequence length and the longest ORF length. Figure 3 shows the ORF Mapper window where the recommended ORF is highlighted.

The Pattern finder is a tool that helps search for a pattern based in the selected reading frame. In the window, the matched portions of the DNA sequence are turned into a different color for easy spying and the number of the occurrences of the query pattern is also displayed. The patterns frequently searched in the analysis include the initiation codon, termination codon, *att* site, and so on.

The Biostatistics viewer window shows some statistical information for the translated sequences. At the current version, the developed system provides the information about the composition and distribution of amino acids for the selected amino acid sequence.

For the later use and further processing, the exporter takes charge of exporting into a file both the DNA sequences collected from GenBank and the translated acid sequences which are confirmed out of 6 possible translations for each DNA sequence. The files can be saved in either the FASTA format or an XML format.

4. Effectiveness Analysis of the Proposed Tool

The aim of the developed tool is to help reduce analysts' intervention in the information gathering, sequence collection and translation incurred biological analysis. We have conducted an experimental work to see the effectiveness in deploying the proposed tool. Suppose the situation that a molecular biologist wants to analyze the HIV virus B subtype for Korean clad. The analyst first wants to gather information about HIV B subtype by retrieving the NCBI nucleotide database, and determines the items which he intends to analyze for the comparative works. For each selected item, he should retrieve the corresponding pages and manually extract out the DNA sequence data along with other related data such as GI number, annotation, and so on. Suppose that 100 items are selected. Since it takes on about 15 seconds to handle each item, the analyst have to spend about minutes to do such bothersome find-cut-and-paste operations. The developed tool does this work on behalf of the analyst.

In order to translate a DNA sequence into its corresponding amino acid sequence, ExPASy's *Translate* can be used in the conventional analysis. When he uses *Translate*, the analyst must copy the DNA sequence and paste it into the input box of *Translate* interface. The translated results should be manually copied into a file. This kind of translation work should be for each DNA sequence one by one. This manual task would take 25 minutes for 100 DNA sequences. The developed system performs this task without the analyst's intervention and in addition recommends most likely one among 6 possible translations for each sequence.

In order to validate the translated amino acid sequences, the analyst would run NCBI BLAST for each sequence which gives related registered amino acid information. The response time of BLAST depends on its system load and thus is not fixed. According to the experience, the response times have been around 15 seconds and the housekeeping time to enter the query and copy the result is around 10 seconds. Therefore, to take care of 100 items, it would take around 41 minutes. The task can be done by the developed system without the analyst's intervention.

As we have seen above, the analyst has to spend about 1 hour 31 minutes to deal with 100 items when he uses the databases and tools in a conventional way. The developed tool allows the analyst to feel free from this bothersome housekeeping task because it performs the chores in an autonomous way without the analyst's involvement.

5. Conclusions

A role of bioinformatics tools is to reduce the analysis burden by automating time-consuming and bothersome manual manipulation works and thus to increase the productivity in the analysis work. Many molecular biological studies ask to collect a volume of DNA sequences from the public databases and to transform into amino acid sequences as a prerequisite for further studies.

This paper presented a system designed and implemented to meet the aforementioned needs. The developed system is capable of collecting DNA sequences at one with the help of a web robot and recommending the most likely amino acid sequences by taking into account the longest ORFs across the possible translations.

For further studies, there remains to add on more functionality to the auxiliary services like regular expression query support for the Pattern Finder, additional biostatistical information provision for the Biostatistics Viewer. We are working on the functionality implementation to collect multiple DNA sequences which match the query sequence to some extent from the public biological database with the help of the robot agent at a time and to choose the sequences matched with the descriptive query, i.e., of which annotation part or title part is compatible with the query statement with the consideration of Gene Ontology information.

6. References

[1] P. Baldi, S. Brunak, *Bioinformatics : The Machine Learning Approach* (2nd Ed.), The MIT Press, 2001.

[2] NCBI National Center for Biotechnology Information, http://www.ncbi.nih.gov/.

[3] NCBI GeneBank, http://www.ncbi.nih.gov/Genbank/index.html.

[4] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, Ron D. Appel and A. Bairoch ExPASy, : the proteomics server for in-depth protein knowledge and analysis, *Nucleic Acids Research*, 2003, Vol. 31, No. 13, pp. 3784-3788.

[5] UniProtKB/Swiss-Prot, http://www.ebi.ac.uk/ swissprot/.

[6] A. Marchler-Bauer, J. B. Anderson, et al., CDD: a curated Entrez database of conserved domain alignments, *Nucleic Acids Research*, 2003, Vol. 31, No. 1, pp.383-387.

[7] G. B. Fogel, D. W. Corne, *Evolutionary Computation in Bioinformatics*, Morgan Kaufmann Publishers, 2003. .

[8] T. A. Brown, *Genomes* (2nd ed), Oxford, United Kingdom: Wiley-Liss, 2002.

[9] S. Aluru, *Handbook of Computational Molecular Biology*(Eds.), Chapman & Hall/CRC, 2006.

저 자 소 개

이혜리
14권 5호 참조

이승희
충북대학교 전자계산학과 석사 과정

이건명
14권 5호 참조

김성수
15권 1호 참조

이찬희
충북대학교 미생물학과 교수

이성덕
충북대학교 정보통학과 교수