지능적 다중염기서열 변환 도구의 설계 및 구현

Design and Implementation of an Intelligent Multiple DNA Sequence Translation Tool

이혜리[1], 이건명[1], 이찬희[2], 이성덕[3], 김성수[1]

[1] 충북대학교 전기전자컴퓨터 공학부
[2] 충북대학교 생물학과, [3] 충북대학교 통계학과
E-mail: winerose@ailab.cbnu.ac.kr

요 약

계통분석을 하는 생물학자들은 관련된 분석대상에 대한 정보를 확보하여 비교분석하기 위해 NCBI 등으로부터 염기서열을 확보하여 아미노산 서열로 변환하는 작업을 수행하게 된다. 많은 서열 데이터에 대해서 데이터베이스로부터 데이터를 검색하고 이를 변환하는 작업을 순차적으로 분석자가 관여하여 작업하는 것이 현재 분석환경이다. 따라서 본 논문에서는 분석의 효율성을 향상시키기 위해, 관심서열의 등록번호(Accession Number) 리스트를 입력하면 해당 서열에 대한 정보를 NCBI로부터 웹로봇을 통해 자동으로 확보한 다음, 확보된 염기서열 전체를 아미노산 서열로 자동 변환하여 가장 긴 ORF(Open Reading Frame)을 추천해주기 위해 설계된 지능형 다중 염기서열 변환 도구에 대해서 소개한다.

Key Words : 생물정보학, 염기서열변환, ORF Mapper

1. 서 론

생물정보학이 추구하는 가장 중요한 목적 중의 하나는 컴퓨터를 이용해 유전체들을 분석하여 이들의 상관관계나 연관성을 파악해 각 유전자들의 연관 정보들을 예측하는 것이다. 그리하여 새로운 서열을 찾아내었을 때 관심 있는 서열들의 유사성과 차이점을 분석해서 염기와 아미노산 수준에서 서열간의 구조적, 기능적 및 진화론적 관련성을 추론할 수 있다[1]. 이를 위해 대부분의 분석자들은 NCBI [2]에서 제공하는 데이터베이스로부터 염기서열을 확보하고 공개용 또는 상용 애플리케이션이나 웹 브라우저 등을 이용하여 분석을 위한 순차적인 작업을 수행하게 된다.

현재 이와 같은 분석환경에서는 염기서열을 아미노산 서열로 변환하고자하는 경우, 분석자가 NCBI의 GenBank 데이터베이스[3]에서 등록번호(GenBank Accession Number : GB)나 GI (Geninfo Identifier) 번호를 입력하면 해당 염기 서열에 대한 정보를 검색하여 결과를 저장한다. 이렇게 확보된 염기 서열은 Web 상의 ExPASy[4]와 같은 변환 프로그램을 이용하여 분석자에 의해서 하나씩 아미노산으로 변환하여 다시 파일이나 원하는 형태로 일일이 옮겨서 저장하게 된다. 이런 일련의 작업은 하나의 염기 서열을 대상으로 하기 때문에 많은 염기서열을 다루기 위해서는 분석자의 많은 시간과 노력을 요구하고 있다.

본 논문에서는 변환 과정의 효율성을 향상시키기 위해서 이미 저장되어 있는 FASTA 형식의 염기서열 파일뿐만 아니라 GenBank 데이터베이스에 등록번호 리스트 파일을 입력하면 해당하는 염기서열을 웹로봇을 통해 자동으로 확보하는 동시에 변환 도구 내에서 이 서열들을 아미노산 서열로 변환한다. 그리고 각 서열에서 가장 긴 ORF(Open Reading Frame)를 추천하여 유전자의 위치를 검색한다. 이렇게 변환된 아미노산 서열들은 FASTA 형식으로 저장 가능하게 함으로써 차후 관심 있는 분석을 위한 기초 자료를 제공한다.

본 논문은 2006년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었습니다.

2. 관련 연구

분석자들이 생물학적 정보를 얻기 위해 접근하는 대표적인 데이터베이스에는 GenBank, EMBL의 UniProtKB/Swiss-Prot[5], PIR-nternational[6], 등이 있다. 온라인상에서 데이터를 링크(link)를 이용해 데이터에 접근하는 것과 연합 데이터베이스의 이용하는 것이 있다. 데이터 제공되는 형식에는 ASN.1 이나 XML과 같은 표준화된 언어로 데이터를 제공하는 경우가 있고, FASTA 형식으로 제공하

는 경우가 있다.

이런 데이터베이스에서 데이터를 검색할 경우 자신이 원하는 염기 서열을 얻는 것이 쉽지 않다. 이것은 염기 서열의 제목만 보고 검색할 때 제목에 적혀있는 정보가 정확한 이름이 아닌 경우가 많기 때문이다. 가장 정확한 검색 방법은 그 유전자가 처음 보고된 논문을 참조하여 GenBank의 등록번호를 이용하는 것이다. GenBank의 모든 정보들은 고유 등록번호가 붙어있어 등록번호를 입력하면 일일이 처음부터 검색하지 않아도 단번에 찾을 수 있다.

본 논문에서는 서열을 검색하기 위하여 관심있는 염기서열의 등록번호가 적힌 리스트 파일을 Batch Entrez Tools에서 불러와서 서열을 검색한 후 그 결과를 프로그램에서 바로 읽어들인다. Entrez는 NCBI에서 개발된 것으로 NCBI에서 제공하는 모든 데이터베이스에 대해서 구분되었으나 관련 있는 자료들에 대한 검색이 가능하다. Entrez[7] 시스템은 염기서열과 단백질 서열자료들 뿐만 아니라 분자모델 3차구조 (MMDB), 게놈과 map 자료, 그리고 문헌에 대한 접속을 제공한다.

[그림 1]은 서열 검색 결과를 FASTA 형식의 파일로 저장하여 보여준 것이다.

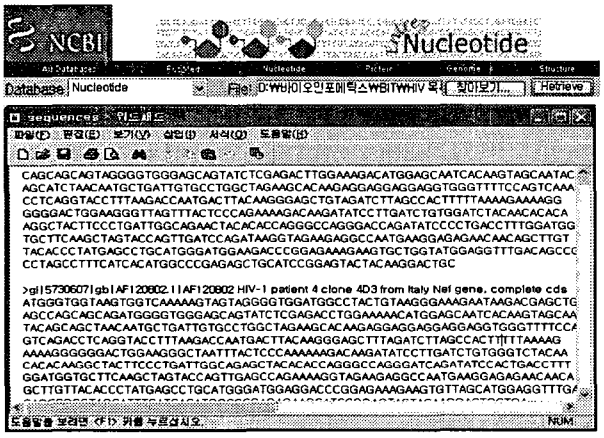

그림 1. 서열정보 검색결과

현재의 아미노산 변환 환경은 확보된 염기 서열을 이용해서 분석자가 Web의 ExPASy (Expert Protein Analysis System)에서 제공하는 있는 여러 가지 Tool 중 translate 부분을 이용해서 아미노산 서열로 변환한다. ExPASy는 SIB(Swiss Institute of Bioinformatics)에서 운영하고 있는 단백질체학 (Proteomics) 전반에 관한 서버를 말한다.

그러나 이것은 다중 염기서열을 한꺼번에 아미노산 서열로 변환하는 것이 아니라 분석자가 수작업으로 서열 하나씩 ExPASy에 복사함으로 변환이 이루어지게 된다. 그러므로 이런 분석 환경은 서열의 개수가 많아지게 될수록 분석자의 어려움이 커진다. [그림 2]는 ExPASy에서 염기서열을 아미노산 서열로 변환하기 전 상태를 보여준 것이다.

본 논문에서 분석자들의 위의 작업환경을 통합하여 일련의 작업을 하는 과정에 사용자와 컴퓨터 사이에 상호작용을 최소화하므로 작업의 효율성을 향상시킬 수 있는 시스템을 제안하고 구현하였다.

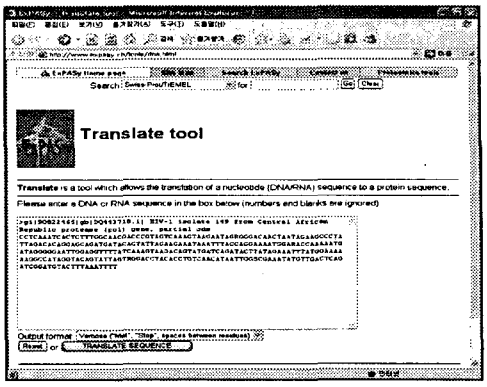

그림 2. ExPASy상의 염기서열

3. 시스템 구성

본 논문에서 설계 구현한 지능적 다중염기서열 변환 도구는 NCBI의 GenBank 데이터베이스 검색 결과를 웹로봇을 이용하여 프로그램 내에서 등록번호별로 자동 분류하고 이를 아미노산 서열로 변환하여 각각의 서열에 대하여 ORF를 검색할 수 있는 기능을 제공하고 있다. 또한 원하는 염기서열 패턴을 아미노산 서열 내에서 검색 가능하고 각 서열의 아미노산 비율을 통계 내며 아미노산 변환 결과를 FASTA 형식로 저장하여 차후 보다 효과적인 아미노산 분석에 편의성을 높이고자 한다. [그림 3]은 전체적인 시스템 구성도를 나타낸 것이다.

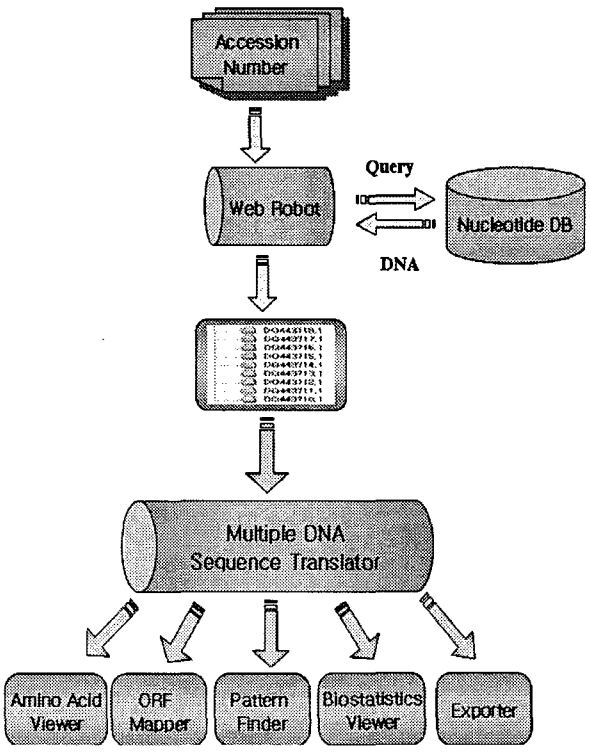

그림 3. 시스템 구성도

4.1. 염기 서열의 확보 방법

관심 있는 염기서열을 확보하기 위하여 먼저 프로그램 안의 메뉴를 통해서 웹브라우저를 연결하여 NCBI 홈페이지의 Entrez에 접속한다. 그 후 등록번호 리스트 파일을 업로드시켜 해당 서열을 검색하고 그 결과를 웹로봇을 이용하여 결과를 받아와서 프로그램 안에서 분류한다.

웹로봇의 동작과정을 살펴보면 먼저 NCBI 의 Nucleotide 데이터베이스에 저장된 등록번호를 검색하기 위한 query를 CGI 프로그램에 추가하면 CGI 프로그램은 입력받은 등록번호를 사용하여 SQL 문장을 생성하게 된다. 웹 로봇은 이 문장을 이용하여 DB에 접속해 해당 염기서열을 검색하고 그 결과를 FASTA 포맷의 형태로 불러와서 프로그램 내에서 등록번호별로 분류한다.

4.2. 아미노산 Viewer

DNA 염기에는 아데닌(A), 티민(T), 구아닌(G), 시토신(C) 4 종류가 있으며 염기 서열 이 바로 유전정보를 나타낸다. 이 4가지 DNA염기쌍 중에서 연속된 3개의 염기가 하나의 아미노산을 지시하는 암호가 되는데 이 형태를 코돈(Codon)이라 한다. 코돈은 유전암호표에 따라 20가지 형태의 아미노산으로 변환되며 각각의 아미노산은 1개의 대문자로 표현되어 아미노산 서열을 이루게 된다.

웹로봇을 통해 확보한 염기서열을 아미노산 서열로 변환할 때 Reading Frame을 선택하여 하나의 염기 서열에서 3가지 다른 형태의 아미노산 서열을 보여준다. Reading Frame이란 염기 3개를 하나의 코돈으로 인식하므로 3개의 염기 중 어느 지점부터 시작해서 해석해야 하는지 읽는 방식을 정하는 것이다. 본 논문에서는 3개의 Reading Frame과 염기서열을 3′에서 5′방향으로 염기서열을 읽어 들여 아미노산으로 변환하는 Reverse Sequence를 선택하게 함으로써 하나의 염기서열에서 총 6개의 서로 다른 아미노산 정보를 얻을 수 있게 하였다. [그림 4]는 웹로봇을 통해 확보한 염기서열을 보여준다. 원하는 염기서열에 대한 Reading Frame을 선택하여 그에 맞는 변환된 아미노산 서열 형태를 보여주고 있다.

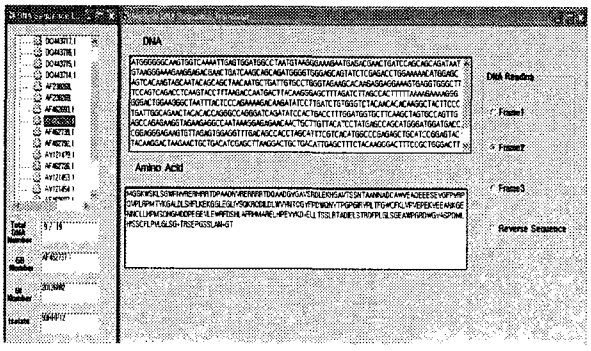

그림 4. 선택한 염기서열에 대한 아미노산 변환

4.3 ORF Mapper

생물학자들은 DNA의 염기서열을 결정하고 아미노산으로 변환한 후, 서열의 어느 부분에 유전자가 있는지 조사하기 위해서 아미노산 서열의 ORF를 검색하는 것이 보편적인 방법이다. ORF(Open Reading Frame)[8]는 각 코돈에 대응하는 아미노산이 개시코돈(ATG)에서 종결코돈(TAA, TGA, TAG)까지의 순차적 배열로, 실제 단백질로 변환되는 부분을 말한다.

생물학에서는 보통 가장 긴 ORF를 찾음으로써 유전자들의 위치를 아는 것이 일반적이다. 그러므로 본 논문에서는 모든 가능한 가장 긴 ORF를 찾기 위해서 M과 *에 대한 모든 가능한 조합을 Frame별로 생성한다. 이렇게 형성된 ORF 중에서 최소 길이(Min Length)값을 두어 그 이상이 되는 값들만 Bar 형태로 표시하여 가장 긴 ORF를 추천하고 그 길이를 나타낸다. 생물학에서 ORF의 길이는 부호화된 단백질의 크기나 분자 무게와 직접 관련되어 있어 추정 ORF에 대한 유용한 척도[9]가 된다.

최소 길이는 변경가능하게 함으로써 ORF 선택에 유연성을 둔다. 또한 ORF를 선택했을 때 전체 아미노산 서열에 대한 해당 아미노산 서열의 위치를 파악할 수 있게 한다. [표 1]은 가장 긴 ORF를 찾기 위한 procedure이며 [그림 5]는 구현한 ORF Mapper의 기능을 보여주고 있다.

Procedure ORF_Mapper

Step 1. Search Every Location of M in Sequence
// M is Initiation Codon
If M Found then
Push Location of M

Step 2. Pop Location of M

Step 3. From the Taken-out Location by Step 2,
Beginning Search *
*// * is Termination Codon*
If * Found Then
Create ORF Object

Step 4. Repeat Step 2 ,3 Until Stack Empty

Step 5. Calculate the Length of Each Object

Step 6. Select the Longest Object

표 1. ORF Mapper의 Procedure

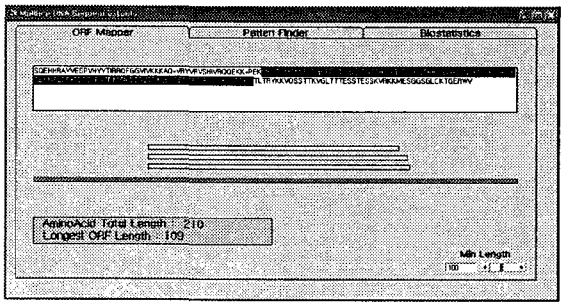

그림 5. 구현한 ORF Mapper

4.4 Pattern Finder

Pattern Finder는 미리 정해진 염기 서열 패턴을 검색하는 것이 아닌 분석자가 입력한 패턴을 Reading Frame 선택에 따라서 검색하는 기능이다. 즉, 찾고자 하는 염기서열 패턴을 입력하면 전체 서열 중에서 입력한 패턴의 위치가 빨간색으로 표시하며 찾은 패턴의 개수를 보여준다. 생물학에서 많이 입력하는 패턴으로 개시코돈과 종결코돈, att site 등이 있다.

4.5 Biostatistics Viewer

아미노산 서열은 단백질의 구조, 기능 및 진화에 대한 정보를 가지고 있다. 본 논문에서는 변환된 아미노산 서열에서 20개의 아미노산이 해당 서열 전체에 대한 비율을 제시하여 차후 단백질 서열 분석과 같은 정보 분석에 유용하게 쓰일 수 있는 기능을 추가하였다. [그림 6]은 해당 서열의 아미노산 구성 분포를 보여주고 있다.

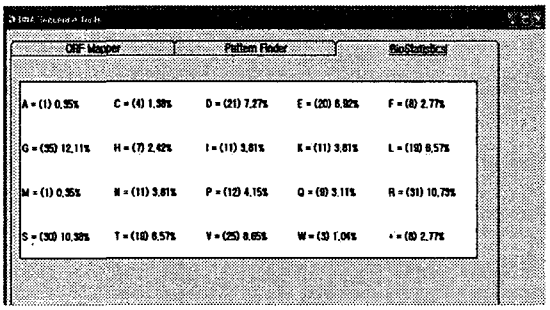

그림 6. Biostatistics Viewer

4.6 Exporter

서열간의 유사성을 검색할 경우 query 서열로 사용할 수 있는 것은 염기나 아미노산 서열이다. 그러나 흔히 진화적으로 멀리 떨어진 서열들을 결과로 얻기 위해 sensitivity를 높이고 싶으면 염기 서열을 번역한 아미노산 서열을 query 서열로 사용하게 된다. 본 논문에서는 분석자의 유사성 검색과 같은 차후 정보 분석에 효율을 높이기 위해서 변환된 아미노산 서열을 FASTA 형식의 파일로 저장한다.

5. 결론 및 향후 계획

생물학자들이 새로운 염기 서열을 찾아내었을 때 기존의 서열들과 유사성과 차이점 등을 분석하여 그 서열이 가지고 있는 정보를 통해 새로운 연관성을 추론하게 된다. 그러므로 염기 서열이 가지고 있는 생물학적인 정보를 분석하는 작업은 매우 중요하다. 지금까지 이러한 작업은 주로 웹이나 PC의 각종 프로그램들을 통하여 이루어지고 있으나 그 일련의 작업이 하나의 프로그램 안에서 동작하기 보다는 서로 다른 웹 사이트와 프로그램을 번갈아 실행시키기 때문에 분석자의 많은 시간과 노력을 요구하는 형태였다.

본 논문은 이러한 과정의 효율성과 기능성 향상을 위하여 지능적 다중 염기서열 변환 시스템을 설계하여 구현하였다. 이 시스템은 웹 로봇을 이용하여 NCBI의 GenBank 데이터베이스에서 등록번호 리스트를 이용하여 원하는 염기서열을 자동으로 확보하고 이 서열을 아미노산으로 변환하고, 변환된 아미노산의 Reading Frame 별로 ORF Mapper를 이용하여 가장 긴 ORF를 추천하였다. 이 때 해당 ORF의 위치를 아미노산 서열에서 찾을 수 있으며 그 길이를 표시하여 정보 분석의 효율성을 높였다. Pattern Finder 기능으로 분석자가 검색 패턴을 입력하면 그 위치와 패턴의 개수를 나타낼 수 있게 하였고 Biostatistics Viewer를 통해 아미노산의 분포를 알 수 있었다. 그리고 차후 단백질 분석과 같은 분석을 위하여 FASTA 형식으로 아미노산 변환 결과를 저장하게 하였다.

차후 이 시스템은 패턴 검색 기능에 Regular Expression 항목과 Match 확률을 포함한 확장된 형태의 Pattern Finder의 기능을 추가하는 등의 현재 제공하고 있는 여러 기능을 좀 더 생물학적 분석에 맞게 보강할 예정이다. 더 나아가 유전자 간의 진화적 관련도 유추하고 기능적으로 혹은 구조적으로 관련된 유전자 그룹에서 공유하는 패턴을 찾아내데 활용할 수 있도록 Progressive Multiple alignment 관한 연구를 계속 할 생각이다.

6. 참고문헌

[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 database and browser : http://www.ncbi.nih.gov/Genbank/index.html

[4] ExPASy Website : http://www.expasy.ch/tools/dna.html

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

[6] PIR-International Protein Sequence Database WebSite : http://mips2.gsf.de/proj/protseqdb/

[7] Entrez Website : NCBI http://www.ncbi.nlm.nih.gov/Entrez

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

[9] " Bioinformatics 기술 및 시장 동향 ", KISTI, 2003.