바이오 데이터 패턴 분석을 위한 시스템 모델링

The Modeling of the System for Bio-Data pattern analysis

송영옥*, 김성영**
우송대학교*, 경북대학교**

Song young-ohk*, Kim sung-young**
Woosong Univ.*, Kyung-Pook National Univ.**

요약

최근 바이오 데이터 분석을 위한 여러 가지 도구들이 있지만 대부분 특정 작업을 위한 작업에 치중되어 있으며 통합된 환경 제시는 아직도 미비한 상태에 있다. 또한 바이오 데이터 분석에 있어 국외의 도구나 데이터베이스에 의존도가 높은 국내 생명과학 분야의 실정을 고려한다면 국산화된 통합 분석환경이 요구되고 있다. 본 논문에서는 바이오 데이터 분석에 필요한 기본 요소들을 모델링하여 효율적인 시스템 개발의 방향제시를 하고자 한다.

I. 서론

생명과학 분야에서 컴퓨터를 활용할 수 있는 대표적인 예로는 서열화, 서열화 분석, 비교, 진화, 돌연변이 추적, 약 설계를 위한 유사성 비교, 단백질 기능 예측, 그리고 세포 메커니즘과 질병 발생에서의 유전자 역할 예측 등 다양한 분야를 들 수 있다. 또한 데이터베이스를 구축함으로써 다른 데이터 연구에서 클로닝 작업을 하고자 할 때 가용성을 제공할 뿐만 아니라 비교 유전학을 위한 기반으로 사용될 수 있다. 바이오 데이터 분석의 가장 초기 과정으로 DNA와 단백질 서열에 대한 데이터 정보 검색을 들을 수 있으며, 이와 같은 생물학적 데이터 마이닝 작업을 하기위해 NCBI, EBI, GenomeNet 등에서 제공하는 데이터베이스를 활용하는 각종 도구들이 출시되고 있다. 이와같은 도구들의 작업으로 정보검색을 위한 스트링 검색과 서열이나 구조의 검색, 배열 및 비교를 위한 유사성 검색과 같은 작업을 수행한다.

이와 같은 기존 분석 도구들의 공통적인 개선점으로는 첫 번째 각각의 분석 작업을 독립적으로 수행하도록 설계되어있다. 이로 인해 분석과정에서 불필요한 작업이 반복되고 있다. 두 번째로는 데이터 분석의 연속성을 고려하지 않았다. 기존 대부분의 도구들이 웹 기반으로 제공되고 있는데 사용자 인증과정을 거치지 않는 단순 request와 response만 이루어지기 때문에 일회성 검색 기능을 제공하기 때문에 같은 결과에 대한 결과값이 다시 요구될 때는 같은 작업을 반복해야 결과 값을 볼 수 있다.

본 논문에서는 이와 같은 문제점을 개선하여 바이오 데이터 분석 기능에서 주요 기능들을 빠르고 연속성 있게 처리할 수 있는 통합시스템의 필요성을 고려하여 이와 같은 시스템 구현에 필요한 전반적인 모델링을 함으로써 시스템 구현의 설계도로 이용하고자 한다.

Ⅱ. 연구배경

바이오인포메틱스의 분야를 크게 3개의 분야로 나누고 있다. 첫 번째는 방대한 생명과학자료를 분석하기 위한 정보 처리 알고리즘과 통계 이론을 개발하는 분야이고, 두 번째는 다양한 형태의 정보와 분석 이론들을 도구화해서 생물학자가 사용할 수 있도록 구현하는 분

야이며 마지막으로 세 번째는 유전자와 단백질의 서열과 구조 및 세포와 개체의 기능에 관련된 방대한 정보를 분석하고 해석하여 생물학적 의미를 찾아내는 분야이다. 첫 번째와 두 번째의 분야에서 컴퓨터공학적으로 접근할 수 있는 요소들이 주로 존재함으로써 컴퓨터공학자들이 생명공학과의 융합을 찾아볼 수 있는 분야라고 할 수 있다. 이에 비해 세 번째 분야는 생물학자들이 접근하는 분야라고 할 수 있다.

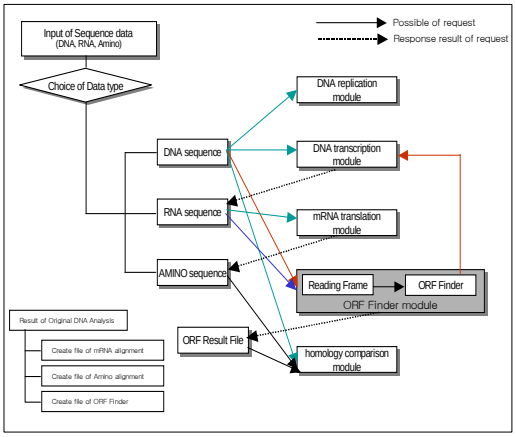
현재까지 일반적으로 이용되고 있는 바이오 데이터 분석 시스템으로는 미국의 NCBI에서 제공되는 데이터베이스를 비롯하여 각종 분석 도구들이 있으며, EBI, GenomeNet등에서 웹 기반 시스템으로 제공되고 있는 분석 시스템을 들 수 있다. 여러 가지 기능을 통합된 형태로 제공되는 분석 시스템으로 대표적인 것은 NCBI, GeneWeb Ⅱ등을 들 수 있다.

본 논문에서는 ORF 검색, 유전자 탐색, 서열 유사성 검색 등의 작업을 고려하여 기존 분석 시스템들의 대표적인 특징 및 개선점들을 조사한 것을 바탕으로 효율적인 바이오 데이터의 분석도구를 구현할 수 있도록 시스템 모델링을 하고자 한다. 본 논문에서는 UML(Unified Modeling Language)을 이용하여 시스템 모델링을 한다.

Ⅲ. 시스템 모델링

1. 분석 시스템 구조 설계

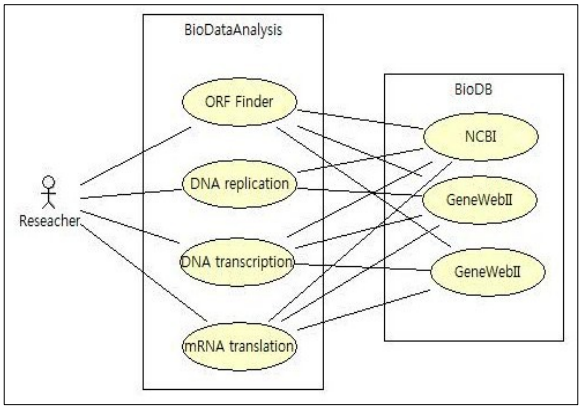
바이오데이터 분석도구에는 다음 〈그림 1〉과 같은 기능들이 추가될 것이다.



▶▶ 그림 1. 전체 시스템 구성 요소

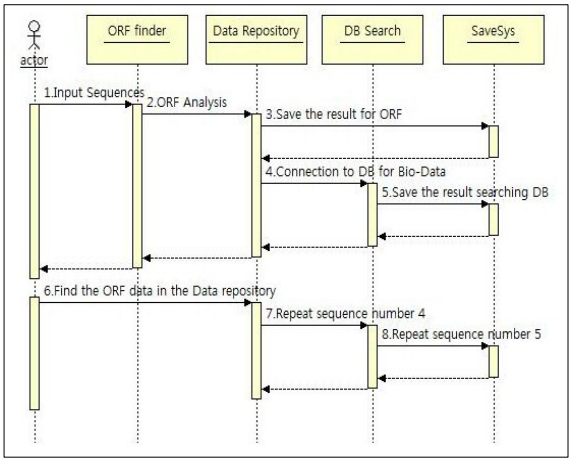
2. UML 모델링

본 논문에서 설계 모델로 제시한 전체 구조를 조망할 수 있도록 UML 다이어그램으로 모델링하였다. 먼저 다음 〈그림 2〉에서는 본 시스템에서 제공하고자 하는 주요 요소를 유스케이스 다이어그램으로 표현하였다.



▶▶ 그림 2. 유스케이스 다이어그램

〈그림 3〉에서는 주요 시스템 요소 사이의 진행과정을 표현하기 위해 시퀀스 다이어그램을 통하여 절차를 표현하였다.



▶▶ 그림 3. 시퀀스 다이어그램

Ⅳ. 기대효과 및 향후 방향

분석 시스템에 포함될 데이터 분석 요소로는 기본적인 작업과정인 DNA전사, mRNA 번역, ORF 검색등과

이러한 작업과정을 통해 얻어진 결과들을 이용하여 데이터베이스로부터 서열 유사성 비교를 하고 전체 유전체로부터 유용한 유전자 탐색 등의 과정이 포함될 것이다.

기존 시스템의 개선점으로 제시하였던 분석 작업의 연속성, 수동적인 입력 과정의 과다로 인한 오류 발생률이 증가할 수 밖에 없었던 사항에 대해서는 본 시스템에서 크게 개선될 것으로 예측한다.

■ 참 고 문 헌 ■

[1] Cynthia Gibas & Per Jambeck, Developing Bioinformatics Computer Skills, 2001

[2] Andreas D.Baxevanis & B.F.Francis Ouellette, Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 2000.

[3] James Tisdall, Beginning Perl for Bioinformatics, 2001

[4]http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

[5]http://www.genome.ad.jp/dbget/db_growth.html