기능 유전체학을 지원하는 유전자 서열 분석 및 관리시스템

허진석* · 김현식 · 진훈 · 김인철**

Gene sequence analysis and management system for supporting functional genomics

Jin-Seok Heo* · Hyun-Sik Kim · Hoon Jin · In-Cheol Kim

요약

본 논문에서는 하나의 시스템 안에서 효율적인 유전자 데이터의 관리와 다양한 서열 분석작업이 가능한 기능 유전체학을 지원하는 서열 분석 및 관리 시스템인 GWB(Gene WorkBench)를 설계하고 구현하였다. GWB는 로컬 데이터베이스 관리뿐만 아니라 GenBank, EMBL, SWISSPROT와 같은 외부 공공 데이터베이스에 대한 접근 기능도 제공하며, 권한을 가진 내부 이용자와 그렇지 못한 외부 이용자들을 구분하여 일부 유용한 기능들은 외부 사용자들도 이용할 수 있도록 설계 되었다. 또 GWB는 유전자에 관한 문헌정보 검색과 관련 유전자 탐색 기능 등 일부 유전자 기능 연구를 지원하는 기능을 제공하고 있다.

Key words : 기능 유전체학(Functional genomics), 유전자(Gene), 서열(Sequence)

1. 서론

휴먼게놈 프로젝트를 계기로 급속한 발전을 보이고 있는 생명과학연구들로 인해 유전자를 비롯한 생명체에 관련된 정보량이 급속하게 증가하게 되었으며, 이러한 다량의 정보를 효과적으로 분석하고 관리하기 위한 생물정보학(Bioinformatics) 기술에 대한 관심도 증가하였다. 또한 인간을 비롯해 많은 생명체의 전체 또는 일부의 유전자 염기서열이 밝혀졌으며, 이러한 유전자 염기서열을 바탕으로 각 유전자의 구조와 기능을 밝히는 구조/기능 유전체학(Structural/Functional Genomics)과 개체간의 차이, 환경에 대한 차이 등을 연구하는 비교 유전체학(Comparative Genomics)에 관한 연구도 활발히 진행되고 있다. 이러한 일련의 연구 활동은 유전학자, 또는 생물학자만으로는 많은 시간과 노력이 필요하다. 이러한 일련의 상황을 생물정보학으로 해결할 수 있게 되었다. 이러한 관심과 함께 생물정보학에 관련된 많은 프로그램들이 개발되고 현재 연구중이지만 아직까지는 생물학과 전산학사이의 많은 차이점이 존재하기 때문에 어떠한 면에서 생물학자의 요구사항을 충족 시켜주지 못하고 있다. 기존의 생명과학 실험실에서의 데이터 관리는 정확한 표준안이 마련되어 있지 않은 상태에서 실험실 내부의 일정한 양식을 정하여 그 양식에 맞도록 작성된 텍스트 파일형태로 이루어져 왔기 때문에 효율적인 서열 데이터의 관리가 이루어지지 않았다. 또한 서열분석에 사용되는 분석프로그램의 종류도 매우 다양하며, 처리 데이터 양식이나 수행환경에 따라 차이가 있고 프로그램간의 연동도 되지 않아 분석 작업이 어렵고 불편하였다. 이러한 문제점들을 해결하기위해서 유전자 서열 데이터를 다루는 일반적인 생명과학 실험실에서 자체적으로 유전자 서열 데이터를 저장 관리하면서, 다양한 서열 분석 작업을 수행하며, 또한 기존의 서열 형식과 상호변환이 될 수 있는 분석 및 관리 시스템이 필요하다. 본 논문에서는 실험실 단위 유전자 서열 분석 및 관리 도구를 설계 및 구현하여 소규모 실험실의 연구 활동을 돕고자 한다. 우선 기존의 유전자 서열 데이터의 관리는 소규모 실험실에 적합한 형태가 되어야한다. 기존의 서열 데이터 관리 방식은 텍스트 파일의 형태를 가지고 관리되었다. 이러한 방법은 효율적인 방법이 아니며, 또한 서열 분석과 연동이 어렵다는 단점을 가지고 있다. 이러한 문제점을 해결하기 위하여 관계형 데이터베이스를 사용하고 또한 실험일지와 서열 파일을 분석하여 실험실

* 경기대학교 전자계산학과
** 경기대학교 전자계산학과

에 적합한 형태의 데이터베이스 스키마를 작성하여 서열 데이터와 사용자 관리가 가능하여야 하며, 또한 서열 데이터의 분석 기능이 포함되어야 한다. 즉 단일 시스템에서 서열 관리 및 분석이 가능한 시스템이어야 한다. 또한 기존 서열 데이터베이스의 형식과 상호변환이 가능해야하며, 아직까지 생물학 데이터에 대한 표준이 마련되지 않았기 때문에 생물학 데이터의 교환이나 이동을 위해서는 XML을 이용하여한다. 이러한 이유로 인하여 XML로 변환하는 기능이 있어야 한다. 본 논문에서 구현하려고 하는 유전자 서열 분석 및 관리 시스템인 GWB(Gene WorkBench)는 로컬데이터베이스 관리뿐만 아니라 GenBank, EMBL, SWISSPROT와 같은 외부 공공 데이터베이스에 대한 접근 기능도 제공한다.

2. 관련연구

2.1 생물정보학

생물체가 지닌 모든 유전 정보의 특성을 분석/규명하려는 유전체 프로젝트의 결과로 방대한 유전체 정보가 생겨나고 있다. 이러한 정보로부터 유용한 정보를 체계적으로 발굴/가공하고 분석하기위해서 생물정보학(Bioinformatics)이 생겨난 연구분야이다. 생물학과 전산학의 경계에 있는 연구분야로 일반적으로 데이터베이스, 알고리즘, 기계학습 및 컴퓨터 그래픽스 등과 같은 컴퓨터 기술을 이용해 생물학 데이터를 저장, 분석 및 해석하는 계산적 생물학을 의미하기도 하며, 생물 시스템의 정보처리 원리를 기초로 컴퓨터나 인공지능 시스템을 개발하는 연구분야로도 인식되고 있다. 또한 이와 관련된 학문으로 다음과 같은 것들이 있다. 유전체학(Genomics), 단백체를 연구하는 단백체학(Proteomics), 기능유전체학(Functional genomics), 구조 유전체학(structural genomics)등의 생물학 관련 연구 분야를 포함하고 있다.

유전체학은 개개의 유전자 수준이 아닌 유전체 전체를 대상으로 연구하는 학문 분야로서, 유전체의 구조와 기능, 진화 등을 다룬다. 또한 생명체가 임의의 환경에서 발현하는 모든 유전자들을 총체적으로 분석하는 분야도 포함한다. 흔히 아래와 같은 분야로 나누기도 한다. 구조 유전체학(structural genomics), 기능 유전체학(functional genomics), 비교 유전체학(Comparative genomics)으로 나눌 수 있다.

비교 유전체학은 말 그대로 각 유전자의 차이를 조사하는 학문으로 특히, 사람간의 유전자 차이를 조사하는 단일 염기 변이는 유전병을 발견하는 중요한 시발점이 되고 있다고 한다. 또한 비교 유전체학을 통해서 각 환자들에게 가장 잘 맞는 약을 투약할 수 있고 이로 인해 치료에도 도움이 될 뿐 아니라, 의료비 절감, 부작용 방지 등 많은 이점들이 있다고 한다.

90년대 까지만 하더라도 세포기능연구에 대한 분자생물학적인 접근은 대부분 단일 유전자나 그 유전자가 발현하는 mRNA의 발현조절을 중심으로 이루어진 반면에, 최근에는 유전체나 단백체 중심으로 총체적생물학 (Hollistic Biology)의 개념으로 바뀌고 있는 추세이다. 기능유전체학(Functional genomics)의 핵심은 개별 유전자의 기능을 다원적으로 시스템을 의미하는 것 같다. 가령, Genomics에서는 DNA 수준에서 주로 유전자의 발현양상을 분석하는 것으로, EST (Expressed Sequence Tag)를 대량으로 chip에 붙이거나 하여 특정 질환이나 상태, 또는 조직이나 species의 표적 유전자가 어떻게 발현하는지를 조사한다 (DNA microarray technology). 프로테오믹스에서는 특정 유전자의 단백질 발현양상과 상호 결합 등을 먼저 규명하여 이 유전자의 정체를 규명하는 것으로 순서적으로는 genomics 와 대별되나 궁극적인 목표는 같은 것이다.

구조 유전체학은 기능을 전혀 모르는 단백질들의 구조 규명을 먼저하고, 기능이 알려진 유사 분자들과의 구조 비교를 하고, 미지의 단백질 기능을 유추한 뒤, 생화학적, 생물학적 검증 실험을 통해 분자의 기능을 규명한다. 다시말해 구조 유전체학은 세포내에 존재하는 단백질들의 3차원적 입체구조를 규명하고 이로 부터 분자와 세포 기능을 유추해 내려는 포스트 지놈시대의 새로운 연구 방향을 제시한다.

단백체학(Proteomics)은 생명체의 전체 유전자, 즉 유전체(genome)에 의해 발현되는 모든 단백질들의 총합을 일컫는 단백체(proteome)를 다루는 학문으로, 어떤 단백질이, 얼마의 양으로, 어떤 환경에서 발현되는 가를 파악하는 것을 목적으로 한다. 생명체의 유전체가 모든 세포에서 동일한 형태로 존재하며, 생명체가 수행하는 기능의 이론적인 면만을 제시할 수 있음에 반해, 단백체는 세포가 처해 있는 환경에 따라, 그리고 고등 생명체의 경우에는 각 조직 별로 유동적으로 존재하며, 세포의 실제적인 기능을 표현해 준다. 이러한 이유로 급속한 속도로 밝혀지고 있는 미지의 유전자들의 기능을 밝혀 내고자 하는 기능 유전체의 한 부분으로 새롭게 부각되고 있고, 세포 내에서 일어나는 실제적인 현상들을 전체 단백질 단계에서 통합적으로 파악하는 수단을 제공한다.

다른 연구분야로 DNA chip이 있다. DNA chip은 수 ㎠ 정도의 좁은 면적에 수백개에서 수만개정도의 올리고뉴클레오타이드나 cDNA를 고밀도로 올려놓고 sample DNA와 교접(hybridization)시킴으로 유전자의 발현(expression)에 대해 알아보거나 SNPs(single nucleotide polymorphisms)을 찾는 용도로 쓰인다. DNA chip은 붙이는 유전물의 크기에 따라 올리고뉴클레오타이드와 cDNA chip으로 구분할 수 있다. cDNA chip에는 최소한 500bp이상의 유전자가 붙여져 있고 올리고뉴클레오타이드 chip에는 15~25개의 염기들로 이루어진 올리고뉴클레오타이드가 놓여있다. 각 chip에 장단점이 있는데 우선 올리고뉴클레오타이드 chip은 정확한 염기서열을 알고 있기 때문에 하나의 염기 변화에 의한 다양성에 대한 연구(SNPs, single nucleotide polymorphisms)도 가능하다. 그러나, 실제로 몇 kb나 되는 유전자에서 극히 일부인 25개 정도의 염기만을 살피는 것이므로 어느 부분을 선택하느냐가 문제가 된다. cDNA chip의 경우에는 염기 자체가

크기 때문에 전반적인 유전자 발현 연구에 쓰는데는 문제가 없지만 SNPs을 볼 수는 없다. 특히 cDNA chip은 두 가지의 다른 환경에서 유전자의 발현정도를 비교하는데 유리하다.

2.2 생물정보 데이터베이스

생물정보학에서 사용할 수 있는 데이터베이스는 개인이나 작은 연구소에서 운영하기에는 비용, 시간, 인력, 정보의 정확성 유지등 많은 제약이 따르기 때문에 대부분의 데이터베이스는 정부에서 주도적으로 운영을 하거나 몇 개의 국가에서 연합을 하여 운영하고 있다. 현재 이러한 데이터베이스를 운영하고 있는 곳은 미국의 GenBank, 일본의 DDBJ, 유럽의 EMBL등이 있다.

<표 1> 생물정보 데이터 베이스

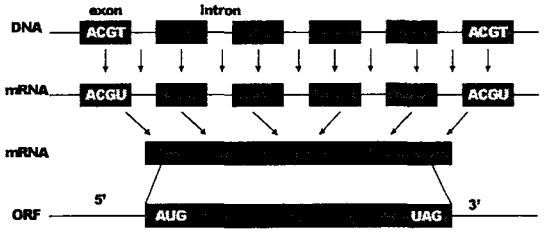
유 형	데이터베이스
DNA, RNA	NeEMBL, GenBank, DDBJ
단백질	Swissprot, PIR, TREML
3차구조	PDB,NAD
유전체	GDB, OMIM
분류학	GENBANK의 Taxonomy database
패턴	Prosite, Blocks, TFD
계통	Gene families, pfam
발현 패턴	Fly view
문헌	Medline, Current Contents
Microarray	NHGRI, Stanford Genome Resource
대사경로	ExPASY-Biochemical Pathway, SWISS-2D PAGE

생물정보학에서 사용할 수 있는 데이터베이스의 범위는 다른 분야에 비해 훨씬 광범위하다. 데이터베이스의 분류하면 크게 다음과 같이 분류할 수 있다. 첫째, DNA와 RNA의 기본이 되는 뉴클레오타이드(Nucleotide)의 정보를 가지고 있는 데이터베이스이다. 이것은 일반적으로 한 종류의 서열 데이터로 전문화된다. 둘째, 단백질(Protein) 서열정보를 저장하고 있는 데이터베이스, 셋째 단백질의 3차원 구조정보를 저장하고 있는 데이터베이스, 넷째 유전체와 염색체의 맵핑에 관한 정보를 가지고 있는 데이터베이스, 다섯째 생물의 분류학적 정보를 가지고 있는 데이터베이스, 마지막으로 생물체에 관련된 문헌 정보를 저장하고 있는 데이터베이스로 구분할 수 있다. 다음 데이터베이스에 대한 설명이다. <표 1>은 데이터베이스의 유형과 실제 운영되는 데이터베이스를 표로 요약한 것이다.

2.3 유전자 서열분석

유전자 탐색은 생물체로부터 얻어진 염기서열은 네 개의 염기 A(아데닌), T(티민), C(사이토신), G(구아닌)로 구성되어 있으며, 이 서열에는 유전에 관여하는 exon과 유전에 관여하지 않는 intron을 구성되어 있다. 이 때 T는 U(우라실)로 치환된다. 이 염기서열에서 intron을 제거한 exon만으로 구성된 mRNA가 생성된다. 유전자 암호인 코돈은 3개의 염기로 구성되며, mRNA 전체가 모두가 단백질로 바뀌는 것이 아니기 때문에 단백질로 바뀔수 있는 부분을 찾아내는 것이다. 이 부분을 ORF(Open Reading Frame)라 하며 ORF의 시작 코돈은 AUG, 종료코돈은 UGA, UAG, UAA이다. [그림 1]은 염기 서열에서 mRNA 생성하여 ORF를 찾아내는 과정이다. 이와 관련된 도구로는 GENSCAN, GRAIL 등이 있다.

[그림 1] 유전자 탐색



유전자통계분석은 염기서열에서 염기의 수, 각 단백질을 의미하는 코돈의 수, 전체 코돈의 수등 서열의 통계적 정보를 제공한다. 서열번역은 유전자 탐색에 의해서 얻어진 DNA 서열을 각 코돈에 대응되는 단백질로 변환하는 과정이다. 이단계에서 사용되는 아미노산의 수는 20개이다. DNA 서열에서 단백질로 번역은 가능하나 단백질에서 DNA 서열로의 역번역은 성립하지 않는다. 유사서열 검색은 DNA 서열이나 단백질 서열을 여러 쌍으로 비교하여 상동성이 존재하는 서열을 찾아내는 것이다. 이것은 두 서열간의 진화적 관련도를 나타낸다. 여기에 사용되는 알고리즘으로 BLAST/FASTA등이 있으며, 알고리즘의 이름으로 도구 또한 개발되었다. [그림 2]은 상동성 검색을 위한 서열 짝 정렬을 나타내고 있다.

[그림 2] 서열 짝 정렬

```
seq1        ----------ATGCTAGCATGCTAGCTGTGGTCTAGTC
seq2        CAGTCGATCGATGCTACGATGCATGCTG----------
```

다중 서열정렬은 3개 이상의 DNA 서열 또는 단백질 서열을 하나의 정렬로 나타내는 것으로, 패밀리 분석, 계통관계분석, 도메인 분석 등의 기능분석 연구를 위해 사용된다. [그림 3]은 다중 서열 정렬을 예를 나타내고 있다. 이와 관련된 알고리즘으로 Clustal, MSA등이 있다.

[그림 3] 다중 서열 정렬

```
seq1        ATGCTAGC-TAGCTAGCTAGCTA-GCTAGCT--
seq3        -GACTAGCACATGCATCTAGCTA-GCTAGCT--
seq2        ----CGATGCAGCATGCTGACGATGCATGCTGA
```

제한효소검색에서 제한 효소는 보통 이중 나선 DNA의 특정한 4-8염기의 서열을 특이적으로 인식하여 그 부위에서 DNA를 절단 시키는 효소이다. 주로 DNA를 재조합하기 위하여 사용된다. 한편, 현재 운영되고 있는 유전자 서열 관련 데이터베이스들은 서열에 필요한 부가 정보들이 서로 다르고 또한 같은 내용이라도 서로 다른 형식으로 표현 되어있기 때문에 상호 간에 변환 과정이 필요하다. 하지만 이러한 경우에 서로 다른 정보를 나타내고 있기 때문에 변환시 정보의 손실을 가져올 수 있다.

3. 설계

시스템을 설계하기에 앞서 실험실에서 하는 연구에 대한 정확한 이해가 필요하며 또한 그 연구에 필요한 요소들은 무엇인지에 대한 정확한 인식이 필요하다. 우선 기존의 서열 분석과 관리가 별도로 존재하던 것을 동시에 제공 할 수 있도록 설계해야 하며 또한 본 논문에서 사용되는 서열의 데이터 형식이 외부의 다른 데이터베이스와 호환될 수 있도록 외부의 서열데이터 형식으로 변환이 가능하며, 이때 정보의 손실이 없어야한다. 또한 분석 기능과 데이터베이스가 효율적으로 연동 될 수 있도록 설계하기 위해서 각 분석 방법들의 특징을 이해해야 한다.

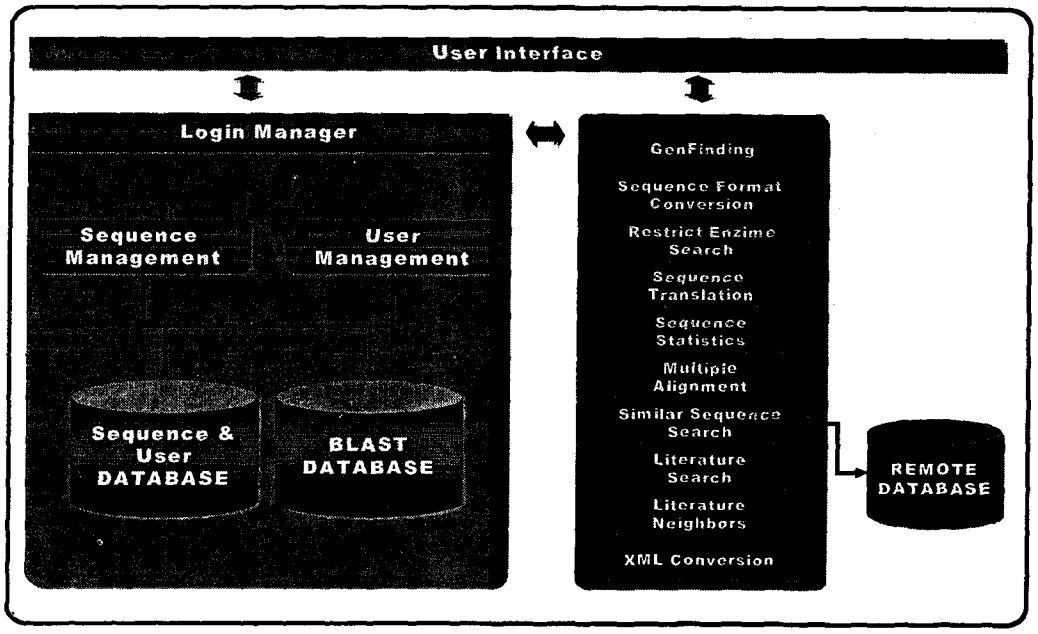
3.1 시스템 구성

시스템은 크게 서열 관리, 사용자 관리, 유전자 서열 분석으로 구성되어 있으며 구조는 [그림 4]와 같다. 사용자는 웹을 이용하여 시스템에 접근할 수 있다. 웹 이용하는 것은 그 만큼 외부에 노출이 되기 쉽기 때문에 보안에 보다 많은 신경을 써야한다. 사용자는 Login Manager를 이용해서 서열 및 사용자 데이터를 보호하게 된다. 데이터베이스에는 Login Manager를 통해서만 접근할 수 있다. 사용자의 권한에 따라 서열을 등록, 검색, 삭제하는 기능을 사용 할 수 있다. 등록절차를 거친 내부 사용자의 데이터를 허가받지 않은 사용자로부터 보호하는 역할을 한다. 이를 위해 GWB에서는 사용자를 관리자, 내부 사용자, 외부 사용자로 구분하였고 로컬 데이터베이스에 구축된 정보들은 외부 사용자가 접근하여 사용할 수 없도록 하였다.

외부 사용자도 서열 분석 기능만을 이용할 경우 일반적으로 사용되는 방식을 따라 서열정보를 복사해서 입력 후 이용할 수도 있고 파일 업로드 과정을 거쳐서 이용할 수도 있다. 서열분석은 이미 밝혀진 서열의 정보를 바탕으로 서열 사이의 유사성 및 이질성을 분석하고 서열수준에서 유전자의 기능을 예측하는 과정이며 관련 연구에서 언급한 기능들을 수행하도록 하였다.

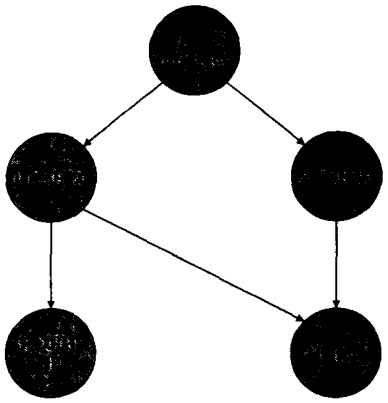
서열관리 부분은 기존에 존재하는 서열 데이터베이스가 색인 순차 파일(ISAM)형태로 관리되고 있던 반면에 이 시스템에 서열 관리는 관계형 데이터베이스를 사용하도록 설계하였다. 이 부분은 사용자가 입력한 서열을 등록, 수정, 삭제하는 기능을 하며, 또한 서열분석 부분과 연동이 되는 부분으로 시스템에서 가장 중심적인 역할을 하는 부분이다.

[그림 4] 시스템 구조



사용자가 신규 서열을 추가하기 사용자는 관리자에게 서열의 등록 요청을 하게 된다. 시스템의 특성상 두 가지의 유형의 데이터베이스를 사용하는데 하나는 관계형 데이터베이스고 또 다른 하나는 유사서열 탐색용 데이터베이스이다. 하지만 이때 사용자가 등록 신청한 서열을 사용자가 직접 입력하게 되면 데이터의 정확성이 떨어질 수 있는 문제점이 있기 때문에 서열을 관리자가 확인한 후 등록하게 된다. 등록된 후에는 사용자가 임의로 데이터를 수정할 수 있으나, 삭제를 할 수는 없다. 또한 서열 관리 부분은 등록된 사용자만이 이용 할 수 있다. 이 부분은 사용자 관리 부분에 인증된 사용자 많이 이용할 수 있는 부분이다.

[그림 5] 사용자별 기능분류



사용자 관리부분은 이 시스템에서 중요한 부분 중의 하나로 데이터베이스에 대한 접근 권한을 관리하는 부분이다. 사용자는 크게 내부사용자와 외부사용자로 구분 할 수 있으며, 또한 내부 사용자는 관리자와 일반 사용자로 나눌 수 있다. 우선 사용자에 따라서 사용할 수 있는 기능을 보면 [그림 5]와 같다. [그림 5]에서 보는바와 같이 내부 사용자는 사용자/서열관리 기능과 서열 분석 기능을 사용할 수 있다. 하지만 외부사용자는 서열 분석 기능만을 사용할 수 있다. 이것은 사용자에 따라서 내부의 데이터베이스의 접근을 제한하고 있기 때문이다. 외부 사용자는 분석 기능을 사용하지만, 서열을 등록, 수정하는 기능들을 사용할 수 없다. 외부사용자가 모든 기능을 사용하기 위해서 사용자 등록요청을 하고, 관리자가 등록처리를 해주면, 그때부터 내부사용자와 같은 기능을 사용할 수 있다.

서열분석은 이미 밝혀진 서열의 정보를 바탕으로 서열 사이의 유사성 및 이질성을 분석하고 서열 수준에서 유전자의 기능을 예측하는 과정이며 관련 연구에서 언급한 기능들을 수행하도록 하였다. 여기에는 이미 많은 훌륭한 프로그램들이 개발되어 사용되고 있으며 성능상으로도 입증된 것들이 많이 있다. 그러므로 우리는 새로운 알고리즘의 개발 보다는 데이터베이스와 연동을 통한 사용자의 편의성에 목적 두었다. 유사 서열 검색을 위해 NCBI에서 제공하는 BLAST[모듈을 이용하며 BLAST의 결과를 파싱하여 결과를 재생성하여 사용자로 하여금 이해를 높이도록 하였다(Altschul, S.F. et. al., 1990.). 유사 서열 검색의 경우에는 다른 분석 방법과 다르게 여러 서열이 검색된다. 이렇기 때문에 사용자 쉽게 서열을 찾을 수 있도록 유사도를 그림의 형태로 표현하고, 또한 각 서열을 나타내는 곳을 이미지 맵을 사용하여 표현하였다. 다중 정렬을 위해 가장 많이 애용되고 있을 뿐더러 기타의 다른 작업과의 연동을 위해서도 많이 사용되는 CLUSTALW를 이용하여 데이터베이스나 웹 인터페이스로부터 서열을 입력받아 다중 정렬을 하도록 설계하였다. 유전자 탐색을 위한 모듈로서 GENSCAN을 사용하였다. 문헌 검색은 유전자 이름이 등장하는 문헌을 검색한다. 이를 통해 NCBI의 PubMed라는 도구를 이용할 수 있다. 또한 PubGene에 대한 질의가 가능하도록 구성하여 본 시스템에서 유사서열들 간의 네트워크를 생성하여 유전자의 기능을 추정할 수 있도록 하였다. 또한 원할한 데이터의 교환을 위하여 XML로 데이터를 변환하는 기능을 가져야한다. 이때 지원하는 XML형식은 BSML (Bioinformatics Sequence Markup Language)과 GAME(Genome Annotation Markup Elements) 형식을 지원하고 있다.

3.2 서열 데이터 형식

GWB에서 사용하는 데이터 형식은 실험실내에서 사용하는데 목적이 있기는 하지만 외부의 다른 서열 데이터베이스와 상호변환이 가능해야 하기 때문에 데이터베이스의 스키마를 구성할 때 주의를 요한다. 기존의 데이터베이스인 GenBank, EMBL, SWISPROT을 서로 다른 형식으로 존재하고 있다. 이러한 형식들은 저마다 나름대로의 장•단점을 갖고 있으며 필요에 따라 다른 형식으로 변환되어 사용되어야 한다.

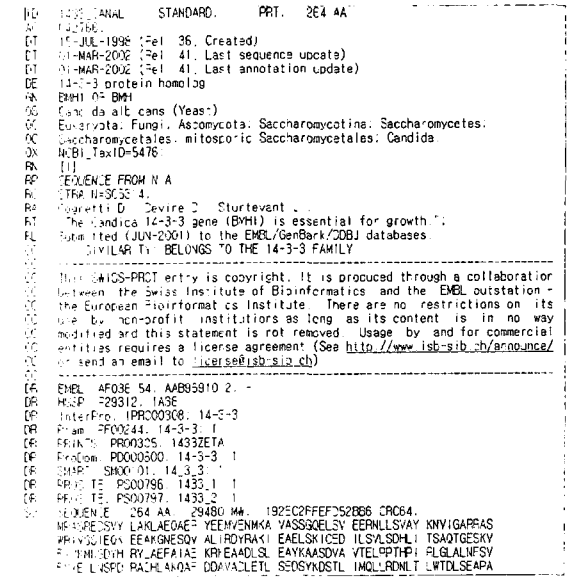
[그림 6] KSF 서열

```
            Tissue Procurement: The Cepko Laboratory
            cDNA Library Preparation: Life Technologies, Inc.
            cDNA Library Arrayed by: The I.M.A.G.E. Consortium (LLNL)
            DNA Sequencing by: Baylor College of Medicine Human Genome
            Sequencing Center
            Center code: BCM-HGSC
            Web site: http://www.hgsc.bcm.tmc.edu/cdna/
            Contact: ama@bcm.tmc.edu
            Gunaratne, P.H., Garcia, A.M., Lu, X., Hulyk, S.W., Hale, S.M.,
            Yoon, V.S., Kowis, C.R., Lawrence, S., Martin, R.G., Muzny, D.M.,
            Richards, S., Gibbs, R.A.
            Clone distribution: MGC clone distribution information can be
            found
            through the I.M.A.G.E. Consortium/LLNL at: http://image.llnl.gov
            Series: IRAK Plate: 61 Row: o Column: 22
            This clone was selected for full length sequencing because it
            passed the following selection criteria: matched mRNA gi: 10946969
            This clone has the following problem: incomplete processing.
FEATURES            Location/Qualifiers

                    /product=""
                    /title=""
                    /value=""
                    /translation=""
                    /modifier="serotype"

                    /product=""
                    /title=""
                    /value=""
                    /translation=""
                    /modifier="serotype"

                    /product=""
                    /title=""
                    /value=""
                    /translation=""
                    /modifier="serotype"
BASE COUNT      713 a    688 c    778 g    656 t
ORIGIN
        1 gcgttccggg agaccgaatt gcggtgtgag ccgccgtggg gactgcgaag cgggcccgga
       61 gtccgagatg gagccgcccg gaggtgcttg aggcgagcgg gcggagcggc tgcgacggac
```

또한 실질적으로 소규모의 실험실에서는 서열에 대한 정보를 위와는 다른 형식으로 훨씬 간단하게 저장하여 관리하고 사용한다. 그러므로 데이터 형

식들 간의 호환성을 제공하도록 구성하는 것이 생물학 정보를 관리하는데 있어서 필요한 기능이라 할 수 있다. GWB에서는 사용하기 적합하면서도 유명 서열 형식과의 호환성을 고려한 새로운 서열 정보 형식(KSF, Kyonggi Sequence Format)을 고안하기로 하였다. KSF는 우선 NCBI의 BankIt에서 요구하는 형식을 수용하여 General Submission/Reference/Source 정보, DNA 서열입력, 추가 정보를 포함하되 소규모의 실험실에 적합하도록 간략화 시켰다. 이것을 우리가 대상으로 하였던 실험실에서 기록되어 오던 기존의 서열정보 형식과 통합시킴으로써 수기로 작성되어 손실이 쉽고 부실하게 기록되어 오던 서열정보를 제대로 관리할 수 있도록 하였다. [그림 6]은 KSF의 데이터 형식을 나타내 그림이며, [그림 7]는 EMBL 데이터 형식을 나타낸 것이다.

[그림 7] EMBL 서열



[표 2]은 KSF의 각 항목을 표로 요약한 것이다. 대부분의 항목은 하나의 입력을 받지만, Feature와 Reference는 여러 개의 항목이 입력 값으로 올 수 있다. 예를 들어 실험에 사용한 서열에서 여러 개의 CDS나 다른 중요한 정보가 발견할 수 있다. 그렇기 때문에 이 부분은 여러 개의 입력을 받을 수 있기 때문에 같은 항목에 여러 개의 추가정보들을 가지게 되다. 이 부분에는 전체 서열에서 그 부분의 시작 위치와 끝 위치, 그 부분의 이름, 단백질로 번역되었을 때의 서열 등을 추가로 기록한다. 또한 Reference도 여러 개의 문헌을 참고할 수 있기 때문에 이 부분 또한 저자, 주제, 제목 등의 추가 정보를 가질 수 있으며, 이 정보 역시 여러 개가 올 수 있다. 이러한 점들은 서열 형식이외에 데이터베이스를 설계할 때와 사용자 인터페이스를 설계할 때에 영향을 주게 된다. KSF는 다른 서열 형식으로 변환되기도 하지만, XML로 변환되기에 적합한 형태로 설계되었다.

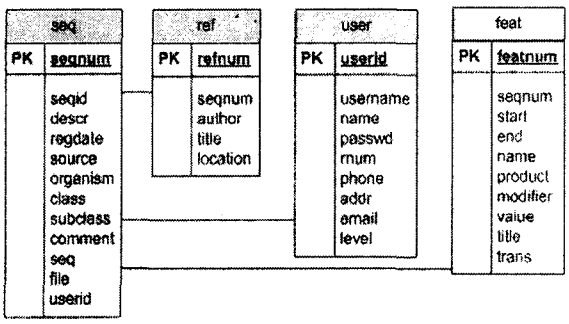
<표 2> KSF의 항목

항목	설명
GeneName	서열의 이름
Definition	서열에 대한 간단한 설명
Source	생물의 일반이름 생물의 공식적 학명과 분류단계별 분류군
Organism	생물의 공식적 학명과 분류단계별 분류군
Classification	원핵, 진핵 결정
Feature	단백질 또는 RNA를 암호화하는 부분에 대한 정보.
Sequence	서열
Description	실험에 대한 설명
Reference	인용문헌(복수가능)

3.3 데이터베이스 설계

본 시스템에서 가장 중요한 부분을 차지하는 부분이 데이터베이스이다. 본 시스템에서는 시스템의 특성상 두 가지의 데이터베이스를 사용하게 된다. 첫 번째는 서열 분석 및 관리, 사용자 관리에 사용되는 관계형 데이터베이스와 유사서열 탐색에 사용되는 색인 순차 파일이다. 색인 순차 파일은 시스템에서 사용하고 있는 BLAST에 종속적인 데이터 형태이다. 이러한 이유로 두 가지의 형태의 데이터베이스가 존재한다.

[그림 8] 관계형 데이터베이스의 테이블 구조

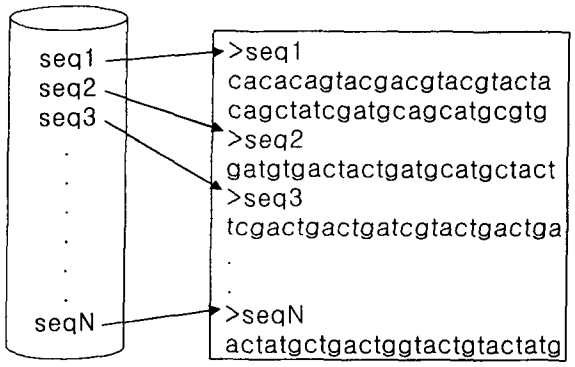


우선 관계형 데이터베이스를 설계하려면 앞에서 언급한 KSF의 특징을 이해해야한다. 우선 데이터의 항목을 살펴보면 크게 입력 값의 수가 정해져 있는 부분과 입력 값의 수가 정해지지 않은 부분으로 나눌 수 있다. 하나의 서열에는 기본적으로 서열과 그 이외의 부가정보가 있으며, 또한 이 부가정보는 입력 값의 수가 일정하지 않다. 그렇기 때문에 순수 서열 정보를 제외한 나머지 Reference/User/Feature정보는 별도의 테이블로 관리하여야 한다. 하나의 서열에는 여러 개의 Reference가 있을 수 있고 또한 한명의 사용자는 여러 개의 서열을 가질 수 있다. 또한 하나의 서열에는 서열의 부분에 관한 특징들이 존재하기 때문에 Feature 정보 또한 별도의 테이블로 존재한다.

seq 테이블에는 KSF 형식의 바이너리 파일이 존재하고 있기 때문에 외부의 다른 형식의 데이터 파일로의 변환이 가능하다.

이 시스템에는 또 하나의 자료구조인 색인순차파일이 존재하는데 이 것은 FASTA형식의 텍스트 파일 형태로 되어있으며 이 것의 인덱스를 만들어 유사서열 검색에서 사용하게 된다. 이 부분은 분석 프로그램과 밀접한 관계를 가지고 있기 때문에 관계형 데이터베이스로 설계하는 것이 불가능하다. 아래의 [그림 9]은 유사서열 검색에 사용되는 색인순차파일을 나타내고 있다.

[그림 9] 유사서열 검색용 색인 순차 파일



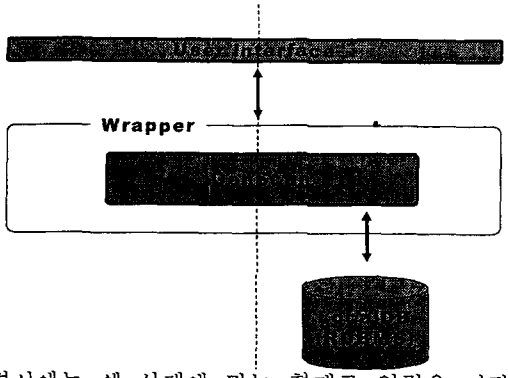
또한 본 시스템은 유사 서열 검색시 두 가지의 데이터베이스를 사용한다. 하나는 시스템 내부에 있는 데이터베이스를 사용하고 또 하나는 외부의 데이터베이스를 사용하는 것이다. 이것을 구분하는 이유는 외부 사용자에게는 내부의 데이터를 보호하고 또한 내부의 사용자에게는 내·외부 데이터베이스를 사용하여 검색의 범위를 넓히기 위해서이다. 외부 데이터베이스는 순차색인 파일 형태로 많은 정보를 포함하고 있다.

3.4 기존 분석 프로그램과의 연동

서열분석(Sequence Analysis)은 독립된 프로그램과 웹 인터페이스를 연동해야 하는 부분들이 있다. 이 프로그램들은 이미 그 능력을 인정받은 프로그램들이고 많은 곳에서 사용하고 있기 때문에 이들의 기능을 구현하는 것 보다는 이 프로그램들을 이용하여 사용자가 쉽게 이용할 수 있도록 하는 것이 보다 나을 것이다. 다음의 세 분야가 기존의 분석 프로그램과 웹이 연동되어야 할 부분이다.

[그림 10]는 유전자 탐색에 이용되고 있는 GenScan과 사용자 인터페이스와 연동을 하기위한 구조를 나타내고 있다. 우선 GenScan은 기존에 유전자 탐색에 사용되는 프로그램으로 웹과 연동시키기 위해서는 GenScan에 입출력을 연결을 해주는 기능이 필요하다. 입출력 기능을 하는 부분이 Wrapper이다. Wrapper는 Bioperl과 Perl을 이용하여 사용자 인터페이스와 GenScan을 연결한다. Wrapper는 쉘 상태에서 실행되는 이 프로그램에

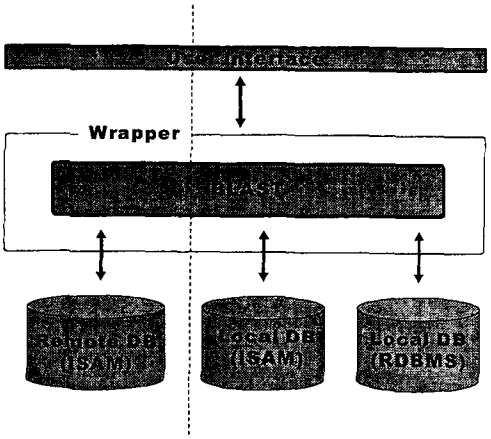
[그림 10] GenScan의 연동



입력시에는 쉘 상태에 맞는 형태로 입력을 전달하고, 결과는 Wrapper가 받아들여 웹에 적합한 형태로 구성하여 사용자 인터페이스에 전달하게 된다. 또한 데이터베이스와의 연결하는 역할을 가지고 있다. 이때 데이터베이스에 접근하기 위해서는 등록된 사용자인지 확인하는 부분이 필요한데 이 부분 역시 Wrapper에서 담당한다.

[그림 11]는 유사 서열 검색에 사용되는 BLAST 프로그램과 사용자 인터페이스를 연동하기 위해서 필요한 구조를 나타내고 있다. Wrapper는 사용자 인터페이스와 데이터베이스를 BLAST와 연동시켜주는 역할을 한다. 자세한 기능을 살펴보면 우선 사용자가 등록된 사용자인지 아닌지를 구분해 주는 기능이 있으며, 이 기능은 사용자의 데이터를 보호하기 위한 기능이다. 또한 등록된 사용자이면 관계형 데이터베이스에서 질의서열을 찾아 BLAST에 입력으로 넘겨주며, BLAST의 결과를 객체로 생성하여 사용자가 이해하고 사용하기 쉬운 형태로 결과를 편집하여 넘겨주는 역할을 한다. 이때 Wrapper는 결과를 그래픽 환경으로 구성하여 사용자에게 보여주어야 한다. BLAST의 여러 가지 파라메터들은 기본값으로 설정하고, 데이터베이스 선택은 사용자 인터페이스에 넘겨받아 BLAST에 전달한다.

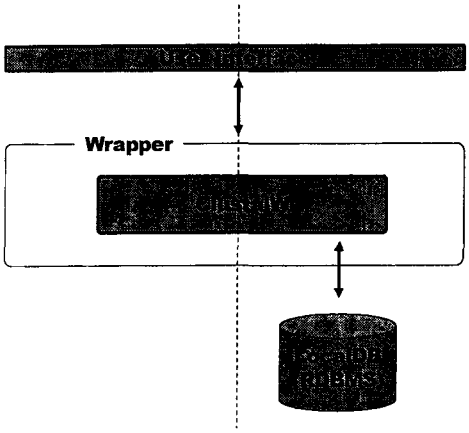
[그림 11] BLAST와 연동



다중 서열 정렬은 ClustalW 사용하여야 하며, 이 또한 사용자 인터페이스와 프로그램간의 연동을 위해서 Bioperl을 이용하여 ClustalW의 출력을 파싱하여 사용자에게 제공한다. 이 부분의 입력은 웹

인터페이스를 통하여 직접 입력하는 경우와 데이터베이스 검색을 통하여 입력하는 부분으로 나누어지며, 외부 사용자는 웹 인터페이스를 통하여 직접 입력하여야 한다. 여기서 중계역할을 하는 부분이 Wrapper이다. 여기서 Wrapper는 사용자 인터페이스, 데이터베이스와 ClustalW의 사이에서 입•출력을 각각에 맞는 형태로 변환하여 각 부분에 전달하는 기능을 한다. 또한 사용자에 따라서 데이터베이스의 사용을 제어한다. [그림 12]은 ClustalW의 연동을 보여주는 것으로 가운데 점선을 사용자의 별 영역을 구분하는 것이며, ClustalW를 감싸고 있는 것이 Wrapper이다.

[그림 12] ClustalW와 연동



4. 구현

4.1 구현환경

본 논문에 구현하고자 하는 GWB 시스템은 기본적으로 웹 환경에 작동하도록 구현하였다. 그리고 본 시스템의 특성상 텍스트처리에 강력한 능력을 가진 프로그래밍 언어를 사용해야 한다. 그렇기 때문에 본 시스템 리눅스 기반의 웹을 이용한 사용자 인터페이스를 가지고 있다. 본 시스템의 구현 환경을 보면 아래와 같다.

<표 3> 구현환경

H/W 환경
- CPU : Intel Xeon 1.8 Ghz × 2
- Memory : 1Gbyte(RDRAM)

S/W 환경
- OS : Linux(Hancom 2.2)
- Language : Perl 5.6.0
- Library : Bioperl 1.0.2, BioXML, GD, Javascript
- External Package : GenSCAN, BLAST,ClustalW
- DBMS : Mysql 3.23.47
- WebServer : Apache 1.3.22

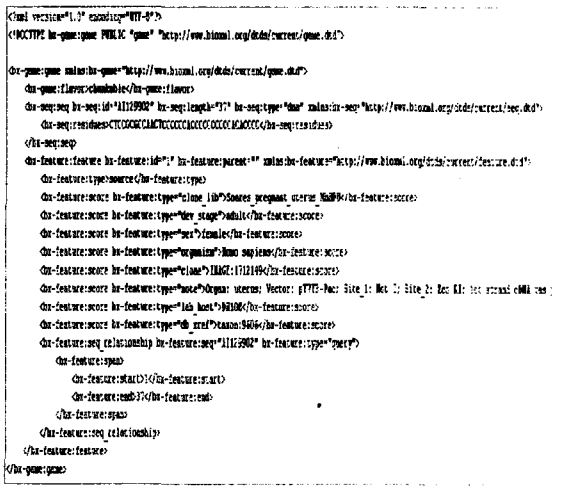
우선 기본적으로 웹 환경에 적합한 언어와 관련된 라이브러리를 사용하였으며, 사용자가 보다 쉽게 사용하고 이해하기 쉽도록 그래픽 라이브러리를 사용하여 사용자 인터페이스를 구현하였다.

4.2 서열 데이터의 관리

이 부분은 기존에 서열 관리 되던 방식을 관계형 데이터베이스인 Mysql과 Perl을 이용하여 구현하였으며 KSF 형식의 서열 파일이 테이블의 한 속성이 되도록 구현하였다. 이 부분은 Bioperl을 이용하여 사용자로부터 입력받은 서열정보를 하나의 객체로 구성하여 데이터베이스 삽입하여 유사서열 검색에 사용되는 데이테베이스로 쉽게 변환 가능하도록 구현하였다. 본 시스템의 주요 기능중의 하나인 서열 데이터의 관리는 앞에서도 이미 설명했지만 등록된 사용자만이 서열을 저장하고 관리할 수 있다. 등록되지 않은 사용자의 경우에는 사용자 등록을 한 후 관리자의 등록하여 주면 GWB의 서열 관리기능을 사용 할 수 있으며 또한 저장된 서열 정보를 바탕으로 다른 분석 작업과 연계하여 사용할 수 있다. 하지만 등록되지 않은 사용자의 경우에는 기본적인 분석 기능을 사용할 수 있기 때문에 분석 기능 간에 연계되지 않는다.

사용자를 구분 하는 이유는 사용자에 정보를 보호하기 위한 것이다. 내부의 데이터베이스에 저장된 정보는 외부의 사용자가 접근 할 수 없기 때문에 사용자의 데이터가 외부에 유출되는 것을 방지할 수 있다.

[그림 13] XML로 변환



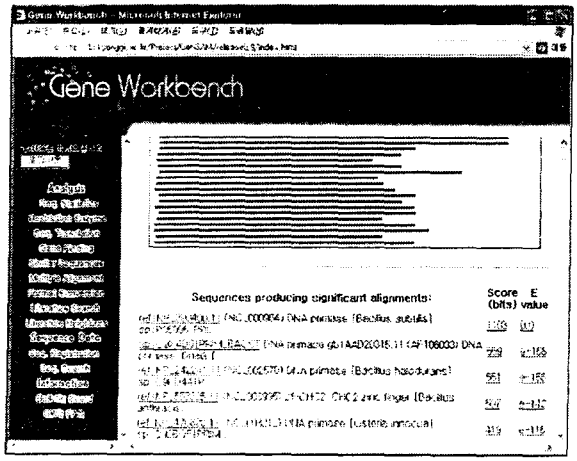
4.3 XML 변환

[그림 13]은 GenBank 서열을 입력하여 XML의 형태도 변환한 화면이다. 이 부분은 Bioperl과 BioXML을 이용하여 구현하였으며, 표준화 되지 않은 서열 데이터의 교환에 있어서 중요한 부분이다.

XML 뿐만 아니라 알려진 대부분의 서열 데이터 형식으로 변환이 가능하다. 입력형식에 맞도록 객체를 생성하여 이 객체를 원하는 형식으로 변환할 수 있도록 구현하였다. 또한 이 부분은 GAME, BSML 두 가지 형태의 XML을 지원한다.

[그림 14]는 BLAST를 이용하여 구현한 부분으로 관계형 데이터베이스와는 별개의 데이터베이스를 구축하여야 한다. 이 부분은 데이터베이스에서 서열의 정보를 얻어 BLAST에 적합한 형태의 데이터베이스를 구축해야 한다. 사용자가 서열을 입력하면, 관계형 데이터베이스 삽입되고 그와 동시에 이 부분에 맞는 데이터베이스에 추가되어진다. 유사 서열 검색은 질의 서열과 유사한 서열을 찾아주는 것으로 질의 서열과 데이터베이스(색인순차파일)에 저장된 서열과의 상동성을 검사하여 가장 상동성이 높은 서열부터 낮은 순으로 정렬하여 사용자에 보여준다.

[그림 14] 유사 서열 검색



5. 결론

지금까지 유전자 서열 분석 및 관리 시스템인 GWB에 대해서 살펴보았다. 기존의 시스템들은 분석 과 관리 기능을 동시에 지원하는 경우가 드물었고, 관리 기능에 일부의 분석기능(서열 검색)등을 지원하고 있었으며, 또한 분석 도구들을 별도로 제공하고 있는 상황이다. GWB에서는 유전자 서열 연구에 있어서 보다 효율적인 시스템을 구현하고자 데이터베이스 관리와 분석 시스템을 통합함으로서 기존 시스템의 단점을 보완하였다.

GWB는 실험실의 일어날 수 있는 상황들을 최대한 고려하여 구현한 시스템으로 서열 데이터의 관리, 사용자 관리, 서열 분석 등의 기능을 가지고 있으며, 또한 서열 분석 프로그램과 데이터베이스의 데이터를 연동하여 보다 쉽게 분석 작업을 수행할 수 있으며, 통합된 분석 시스템으로서 모든 분석 작업을 GWB 내에서 수행할 수 있다. 또한 XML을 지원하여 표준화되지 않은 서열 데이터를 쉽게 교환할 수 있도록 하였다.

또한 GWB는 실험실 내부에서 사용하는 목적으로 구현되었지만, 웹이라는 환경이 가지는 특성을 이용하여 외부의 사용자에게 분석기능을 제공함으로서 실험실 내부뿐만이 아니라 외부의 사용자들도 분석기능을 이용 할 수 있다.

현재의 GWB는 유전자 수준에서의 기능과 일부의 단백질에 관한 기능을 지원하고 있지만 향후에는 단백질 수준의 분석 작업을 지원하는 시스템으로 발전시켜 나가야 할 것이다.

참고문헌

[1] Altschul, S.F. et. al., "Basic local alignment search tool". J. Mol. Biol.215(1990), 403-410.

[2] Altschul, S.F., et. al., "Gapped BLAST and PSI-BLAST:a new generation of protein database search programs". Nucleic Acids Res. 25,(1997) 3389-3402.

[3] Apweiler R., Junker V., Gateau A., O'Donovan C., Lang F., Bairoch A, "New Developments in Linking of Biological Databases and Computer Generation of Annotation: SWISS-PROT and its Computer-annotated Supplement TREMBL", Proc. of the German Conference on Bioinformatics GCB96, Leipzig, Germany 1278(1996), 44-51

[4] Bairoch A., Apweiler R,The SWISS-PROT protein sequence data bank and its new supplement TREMBL, Oxford University Press(1996) 21-25

[5] Burge, C. B. Modeling dependencies in pre-mRNA splicing signals. In Salzberg, S., Searls, D. and Kasif, S., eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam,(1998) 127-163.

[6] Burge, C. B. and Karlin, S. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol.8(1998), 346-354.

[7] Cynthia Gibas, Per Jambeck, Developing Bioinformatics Computer Skills, O'Reilly(2001)

[8] David W. Mount, Bioinformatics : Sequence and Genome Analysis, CSHL

[9] Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30(2002): 2478-83.

[10] Des Higgins, Bioinformatics: Sequence, Structure and Databanks, Oxford University Press(2000)

[11] Gotoh, O., Alignment of three of three biological sequences with an efficient traceback procedure. J. Theor. Biol. 121(1986.), 327-337.

[12] Helen M. Berman, T. N. Bhat, Philip E. Bourne, Zukang Feng, Gary Gilliland, Helge Weissig & John Westbrook The Protein Data Bank and the challenge of structural genomics, Nature Structural Biology, 7 (11)(2000), 957-959.

[13] Henikoff, S., and J.G. Henikoff, Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89(1992.) 10915-10919.