GWB: 유전자 서열 데이터의 관리와 분석을 위한 통합 소프트웨어 시스템

GWB: An Integrated Software System for Managing and Analyzing Genomic Sequences

김 인 철* 진 훈**
In-Cheol Kim Hoon Jin

요 약

본 논문에서는 효율적인 유전자 서열 데이터의 관리와 분석을 위한 웹 기반의 통합 시스템인 GWB(Gene WorkBench)의 설계와 구현에 대해 설명한다. 유전자 서열을 다루는 기존의 시스템들은 서열 데이터의 관리 기능과 분석 기능을 동시에 지원하는 경우가 드물고, 또한 분석 기능 역시 일부 혹은 단일 분석 기능만을 제공하는 단위 프로그램들이 대부분이다. 또 이러한 분석 프로그램들마저 서로 분산되어 있고 다른 수행환경을 필요로 한다. 따라서 이러한 프로그램들을 함께 이용하기 위해서는 많은 수작업과 변환작업을 필요로 하는 등 유전자 서열 데이터를 다루는 많은 생명과학 연구자들이 불편을 겪어왔다. 본 논문에서는 기존 시스템들의 단점을 보완하고 유전자 서열 연구에 효과적으로 도움을 줄 수 있는 보다 편리한 시스템을 구현하고자, 서열 데이터베이스 관리 기능과 다양한 분석 기능들을 하나의 시스템인 GWB로 통합하였다. GWB 시스템 설계의 가장 중요한 이슈는 서로 상이한 분석 프로그램들을 어떻게 하나의 시스템으로 통합할 것이며, 또 이들 프로그램들이 요구하는 서로 다른 서열 데이터 및 서열 데이터베이스 형태를 어떻게 제공할 수 있느냐는 것이다. GWB는 이 문제들을 해결하기 위해 공통의 입출력 인터페이스인 포장기를 이용하여 서로 다른 분석 프로그램들을 시스템에 통합시켰고, 공통 서열 데이터 형식인 KSF를 제안하였으며, 로컬 서열 데이터베이스를 관계형 데이터베이스부분과 색인 순차파일부분으로 나누어 구성하였고, 서로 상이한 서열 데이터 형식간의 변환 기능과 XML 파일로의 변환 기능을 제공하도록 하였다.

Abstract

In this paper, we explain the design and implementation of GWB(Gene WorkBench), which is a web-based, integrated system for efficiently managing and analyzing genomic sequences. Most existing software systems handling genomic sequences rarely provide both managing facilities and analyzing facilities. The analysis programs also tend to be unit programs that include just single or some part of the required functions. Moreover, these programs are widely distributed over Internet and require different execution environments. As lots of manual and conversion works are required for using these programs together, many life science researchers suffer great inconveniences. In order to overcome the problems of existing systems and provide a more convenient one for helping genomic researches in effective ways, this paper integrates both managing facilities and analyzing facilities into a single system called GWB. Most important issues regarding the design of GWB are how to integrate many different analysis programs into a single software system, and how to provide data or databases of different formats required to run these programs. In order to address these issues, GWB integrates different analysis programs byusing common input/output interfaces called wrappers, suggests a common format of genomic sequence data, organizes local databases consisting of a relational database and an indexed

* 정 회 원 : 경기대학교 정보과학부 부교수
kic@kyonggi.ac.kr(제 1저자)
** 정 회 원 : 경기대학교 전자계산학과 박사과정
jinun@kyonggi.ac.kr(공동저자)

sequential file, and provides facilities for converting data among several well-known different formats and exporting local databases into XML files.

☞ Keyword : Genomic Sequence, Sequence Analysis, Wrapper, Integration System, Bioinformatics

1. 서 론

휴먼게놈 프로젝트를 계기로 더욱 활발히 진행되고 있는 생명과학 연구들과 정보기술의 발달로 인해 유전자 데이터를 비롯한 생명 관련 정보의 양이 급속하게 증가하게 되었다. 따라서 이러한 다량의 생명 관련 정보를 효과적으로 분석하고 관리하기 위한 생물정보학(Bioinformatics) 기술에 대한 관심이 더욱 높아졌다. 그리고 그 동안의 많은 유전체 연구들이 주로 인간을 비롯한 다양한 생명체의 유전자 염기서열을 밝히는데 초점이 맞추어져 왔으나 현재는 각 유전자의 정확한 기능이 무엇인지, 그들이 서로 어떤 연관관계를 가지고 있는지 등을 밝히는 기능 유전체학(Functional Genomics)으로 관심이 옮겨가고 있다[6,7]. 이러한 상황에서 유전체 연구에 필수적인 유전자 서열 데이터에 대한 분석 기능과 관리 기능을 제공해 주는 편리한 소프트웨어 시스템에 대한 요구가 더욱 증가하고 있다. 기존의 시스템들은 서열 데이터에 대한 관리 기능과 분석 기능을 동시에 지원하는 경우가 드물고, 또한 분석 기능 역시 일부 혹은 단일 분석 기능만을 제공하는 단위 프로그램들이 대부분이다. 또 이러한 분석 프로그램들마저 분산되어 있고 서로 다른 수행환경을 필요로 한다. 따라서 이러한 프로그램들을 이용하기 위해서는 많은 수작업과 변환작업을 필요로 하는 등 유전자 서열 데이터를 다루는 많은 생명과학 연구자들이 불편을 겪어왔다. 또 기존의 유전자 서열 데이터의 관리시스템들은 정확한 서열 데이터의 표준안이 마련되어 있지 않은 상태에서 실험실 단위의 독자적인 내부 양식에 맞도록 서열 데이터를 저장, 관리해 왔다. 따라서 중요한 유전자 정보의 상호 교환과 공유가 어려웠다. 그리고 대부분 기존 시스템들은 별도의 데이터베이스 관리 시스템(DBMS)의 도움을 받거나 특별한 파일 구성과 색인 방법을 사용함이 없이 단순한 플랫파일(Flat File) 형태로 서열 데이터를 관리하였기 때문에 데이터의 양이 증가할 경우 데이터의 효율적인 이용과 관리가 어려웠다.

이러한 기존 시스템들의 문제점들을 해결하기 위해서 본 논문에서는 다량의 유전자 서열 데이터를 효율적으로 저장 관리할 수 있으면서 다양한 서열 분석 작업을 통합적으로 수행할 수 있는 소프트웨어 시스템인 GWB(Gene WorkBench)를 설계하고 구현하였다. GWB는 GenBank, EMBL, SWISSPROT와 같은 대규모 공공 유전자 서열 데이터베이스들과는 달리, 유전자 서열 데이터를 다루는 실제 실험실 사용자의 편의성을 최대한 고려하여 설계한 시스템이다. GWB는 웹 기반의 사용자 인터페이스를 제공함으로써 기관 내부와 외부의 사용자 모두에게 이용의 편의성을 제공하면서도, 엄격한 사용자 권한 제한 기능을 제공하여 로컬 서열 데이터베이스의 보안성을 높였다. GWB시스템 설계의 가장 중요한 이슈는 서로 상이한 분석 프로그램들을 어떻게 하나의 시스템으로 통합할 것이며, 또 이들 프로그램들이 요구하는 서로 다른 서열 데이터 및 서열 데이터베이스 형태를 제공할 수 있느냐는 것이다. GWB는 이 문제들을 해결하기 위해 공통 서열 데이터 형식인 KSF를 제안하였으며, 로컬 서열 데이터베이스를 관계형 데이터베이스부분과 색인 순차파일 부분으로 나누어 구성하였고, 서로 상이한 서열 데이터 형식간의 변환 기능과 XML 파일로의 변환 기능을 제공하도록 하였다. 이 밖에도 GWB는 최근 관심의 초점이 되고 있는 유전자의 기능 연구를 직접 지원하기 위한 몇 가지 기

능을 추가적으로 제공한다. 즉 GWB는 대표적인 생명과학 및 의료 문헌 데이터베이스인 Medline으로부터 특정 유전자에 관한 문헌 검색 기능을 제공할 뿐 아니라, PubGene시스템과 연계하여 문헌상 연관도가 높은 다른 유전자들을 찾아주는 기능도 제공한다.

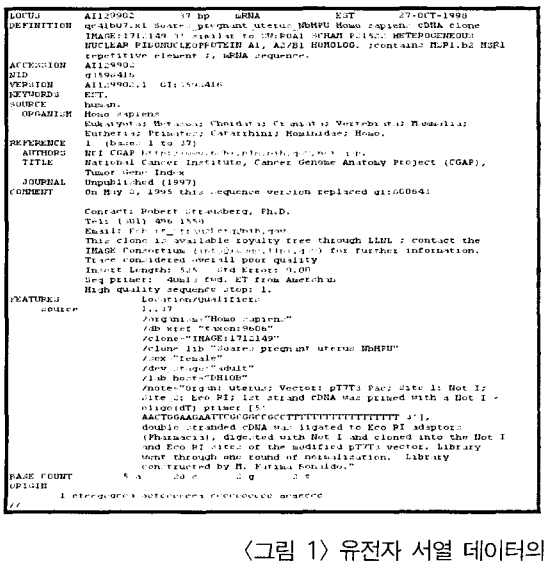
본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련있는 생물정보 데이터베이스들과 유전자 서열 분석방법에 관해 알아보고, 3장에서는 GWB 시스템의 설계에 관해, 4장에서는 GWB 시스템의 구현에 관해 각각 설명한 다음, 끝으로 5장에서 결론을 맺는다.

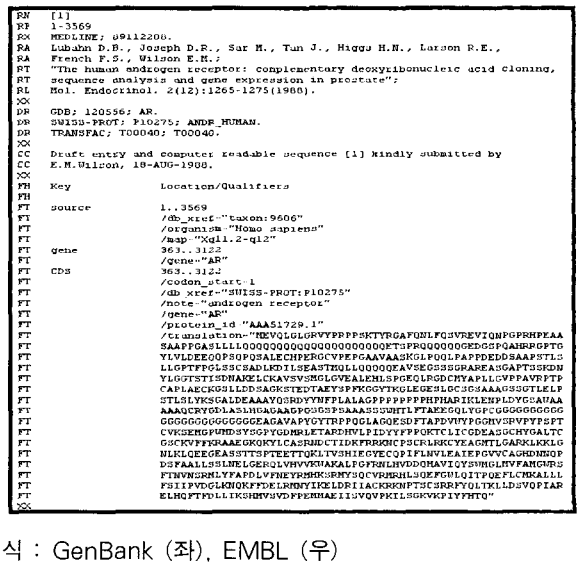
2. 관련 연구

2.1 생물정보 데이터베이스

현재 운영되고 있는 생물정보 데이터베이스는 보유 데이터의 종류와 규모면에서 매우 다양하다. 유형별로는 DNA 및 RNA의 염기(nucleotide) 서열 정보를 가지고 있는 데이터베이스, 단백질(protein)의 아미노산 서열 정보를 저장하고 있는 데이터베이스[3,4], 단백질의 2차 및 3차원 구조 정보를 저장하고 있는 데이터베이스[11], 각 생물의 모든 유전 정보를 포괄적으로 제공하는 유전체 데이터베이스, 단백질 및 각 생물들의 분류학적 정보를 제공하는 데이터베이스, 생물, 의약학 분야의 논문과 문헌 자료를 보유하고 있는 문헌 데이터베이스, 마이크로어레이(microarray)를 이용한 유전자 발현데이터를 저장하고 있는 마이크로어레이 데이터베이스, 단백질들의 대사경로(pathway) 정보를 보유하고 있는 데이터베이스 등으로 구분할 수 있다. 대규모 공공 유전자 서열 데이터베이스의 대표적인 예로는 미국의 GenBank, 일본의 DDBJ, 유럽의 EMBL등을 들 수 있다.

한편, 이들 기존의 유전자 서열 데이터베이스들은 비록 동일한 유전자 서열을 가지고 있다 하더라도 서로 부가 정보 필드의 종류와 길이 등 서열 데이터를 표현하는 방식이 서로 다른 경우가 대부분이다[9]. 예컨대 그림 1은 동일한 유전자에 대한 GenBank 데이터베이스와 EMBL 데이터베이스의 데이터 형식의 차이를 보여준다. 이와 같은 데이터 형식의 이질성은 생물정보의 상호교환과 공유를 가로막는 큰 걸림돌로 작용하고 있다.

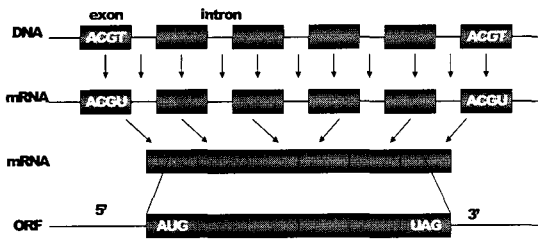




〈그림 1〉 유전자 서열 데이터의 형식 : GenBank (좌), EMBL (우)

2.2 유전자 서열분석

유전자 탐색은 주어진 염기서열로부터 하나의 유전자 정보를 담고 있는 부분을 찾아내는 것을 말한다. 일반적으로 염기서열은 네 개의 염기 A (아데닌), T(티민), C(사이토신), G(구아닌)들로 구성되어 있으며, 이 서열에는 유전에 관여하는 엑손(exon)과 유전에 관여하지 않는 인트론(intron)을 구성되어 있다. 이 염기서열에서 인트론(intron)을 제거한 엑손(exon)만으로 구성된 mRNA가 생성된다. 유전자 암호인 코돈은 3개의 염기로 구성되며, mRNA 전체가 모두가 단백질로 바뀌는 것이 아니기 때문에 유전자 탐색은 단백질로 바뀔수 있는 부분을 찾아내는 것이다. 이 부분을 ORF(Open Reading Frame)라 하며 ORF의 시작 코돈은 AUG, 종료코돈은 UGA, UAG, UAA이다. 그림 2는 염기 서열에서 mRNA 생성하여 ORF를 찾아내는 과정이다. 유전자 탐색을 위한 대표적인 알고리즘과 분석 프로그램들은 GENSCAN, GRAIL등이 있다[5].



〈그림 2〉 유전자 탐색

```
Seq1 - - - - - - - - - - ACGTAGCTAGCTAGCAACTCG
Seq2 - - - - - - ACGATCGAACGTAGCTAGCTA - - - -
```

〈그림 3〉 서열 짝 정렬

```
Seq3 - - - - -ACGTA - - - ACGTAGCTAGCTAGCAACTCG
Seq1 - - - - -ACTGA - - - ACGTAGCTAGCTAGCAACTCG
Seq2 - - - -  ACTGA - ACGAACGTAGCTAGCTA - - - -
```

〈그림 4〉 다중 서열 정렬

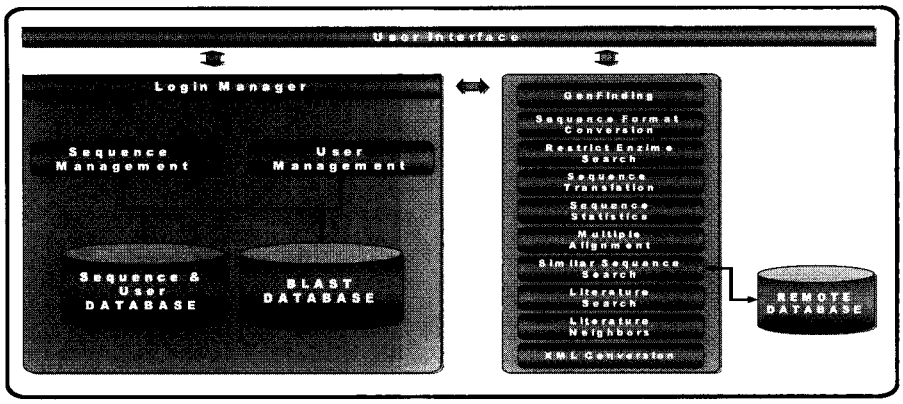
서열 통계 분석은 염기서열에서 염기의 수, 각 단백질을 의미하는 코돈의 수, 전체 코돈의 수등 서열의 통계적 정보를 제공한다. 서열번역은 유전자 탐색에 의해서 얻어진 DNA 서열을 각 코돈에 대응되는 단백질로 변환하는 과정이다. 이단계에서 사용되는 아미노산의 수는 20개이다. DNA 서열에서 단백질로 번역은 가능하나 단백질에서 DNA 서열로의 역번역은 성립하지 않는다. 유사서열 검색은 DNA 서열이나 단백질 서열을 여러 쌍으로 비교하여 상동성이 존재하는 서열을 찾아내는 것이다. 이것은 두 서열간의 진화적 관련도를 나타낸다. 유사 서열 검색을 위한 대표적인 알고리즘과 분석 프로그램들로 BLAST [1, 2], FASTA 등이 있다. 그림 3은 유사 서열 검색을 위한 서열 짝 정렬(Pair-Wise Alignment)을 나타내고 있다.

다중 서열정렬(Multiple Alignment)은 3개 이상의 DNA 서열 또는 단백질 서열을 하나의 정렬로 나타내는 것으로, 패밀리 분석, 계통관계분석, 도메인 분석 등의 기능분석 연구를 위해 사용된다[10]. 그림 4는 다중 서열 정렬을 예를 나타내고 있다. 대표적인 알고리즘과 분석 프로그램들로는 Clustal, MSA등이 있다. 이와 같이 유전자 서열 분석을 위한 기존의 많은 소프트웨어들은 그 기능에 따라 별도의 독립적인 프로그램들로 산재하고 있으며 서로 유기적으로 통합 운영되고 있지 못한 실정이다.

3. 시스템 설계

3.1 시스템 구성

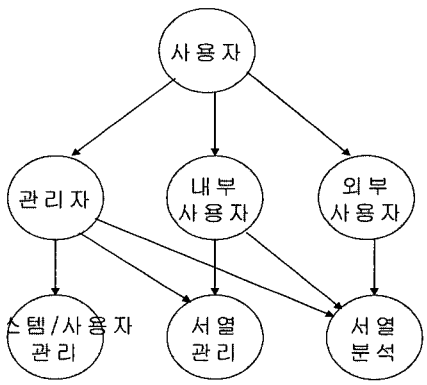
GWB 시스템은 그림 5와 같이 크게 사용자 데이터 관리부, 서열 데이터 관리부, 서열 데이터 분석부 등으로 구성되어 있다. 사용자는 웹을 이용하여 시스템에 접근할 수 있다. 웹 기반의 사용자 인터페이스를 이용하는 것은 사용자 편의성은 뛰어 나지만 그 만큼 데이터가 외부에 노출이 되기 쉽기 때문에 보안에 보다 많은 신경을 써야 한다. GWB 시스템의 로인 관리자(login manag-



〈그림 5〉 시스템 구성

er)는 사용자 관리부의 도움을 받아 인터넷으로 접속하는 사용자들에 대한 인증과정을 수행한다. 따라서 인증과정을 거친 권한을 가진 사용자들만 유전자 서열 데이터의 등록, 검색, 삭제 등을 허용함으로써 서열 및 사용자 데이터를 보호하는 역할을 한다.

GWB의 사용자 데이터 관리부는 이 시스템을 이용하는 사용자들의 기본적인 인적사항을 저장 관리하고, 사용자들의 접근 권한을 관리하는 부분이다. 이를 위해 GWB에서는 사용자들을 관리자, 내부 사용자, 외부 사용자 그룹으로 크게 구분하였고, 그림 6과 같이 이러한 사용자 유형별로 접근 가능한 데이터와 이용 가능한 기능들을 제한하여 운영하고 있다. 그림 6에서 보는 바와 같이 실험실 또는 연구소 등 한 기관의 내부 사용자에게는 서열 데이터를 로컬 데이터베이스에 저장, 관리 가능할 뿐 아니라, 이 로컬 데이터들을 이용한 모든 서열 분석 기능이 가능하다. 하지만 외부 사용자는 서열 분석 기능만을 사용할 수 있다. 즉, 인터넷으로 접속하여 단위 서열 분석 기능을 이용하려는 외부 사용자들에게는 로컬 데이터에 대한 접근은 허용되지 않으나 사용자가 웹으로 직접 제공하는 별도의 서열에 대해서는 유전자 탐색, 서열 정렬, 유사 서열 검색, 서열 번역 등 유용한 대부분의 서열 분석 기능을 제공한다. 외부 사용자도 서열 분석 기능만을 이용할 경우 웹 인터페이스를 통해 서열 데이터를 복사해서 입력 후 이용할 수도 있고 파일 업로드 과정을 거쳐서 이용할 수도 있다. 만약 외부 사용자가 모든 기능을 사용하기 위해서 사용자 등록 및 권한변경 요청을 하고, 관리자가 등록처리를 해주면, 그때부터 내부 사용자와 같은 기능을 사용할 수 있다. 한편 GWB 시스템의 관리자는 내부 사용자에게 허용되는 일반적인 서열 데이터 관리 및 분석 기능 외에도 사용자 권한 제어, 데이터 백업 관리, 색인 관리 등 시스템 관리 목적의 기능을 추가적으로 이용할 수 있다.



〈그림 6〉 사용자별 권한 제어

GWB시스템의 서열 데이터 관리부는 사용자가 입력한 서열을 등록, 검색, 수정, 삭제하는 기능을 수행하며, 또한 저장된 서열 데이터의 분석을 위해 서열 분석부와 연동이 되는 부분으로서 시스템에서 가장 중심적인 역할을 하는 부분이다. 서열 데이터의 체계적인 관리를 위해 신규 데이터의 등록은 관리자의 통제를 받도록 설계하였다. 즉, 일반적으로 한 사용자가 신규 서열데이터를 로컬 데이터베이스에 등록하기 위해서는 데이터를 입력한 뒤 관리자에게 서열 등록 요청을 하여야 하고, 관리자가 이 데이터를 검토한 뒤 등록을 결정함으로써 비로소 정식 등록이 이루어지게 된다. 또 데이터의 보호를 위해 등록한 후에 필요한 경우 등록 사용자가 임의로 데이터를 수정할 수는 있으나, 삭제 할 수는 없도록 설계 하였다.

GWB시스템의 서열 데이터 분석부는 그림 5에서 보듯이 하나의 서열 데이터에 대한 다양한 분석 기능들을 제공한다. 제공되는 대표적인 서열 분석 기능들은 유전자 탐색, 서열 형식 변환, 서열 번역, 서열 통계 분석, 서열 다중 정렬 및 짝 정렬, 유사 서열 검색, 제한효소 탐색, 문헌 검색, 연관 유전자 검색, XML 변환 기능 등이다. 이미 각 분석 기능을 위한 매우 효율적인 단위 분석 알고리즘과 프로그램들이 많이 개발되어 사용되고 있는 실정이므로, 본 시스템에서는 새로운 알고리즘의 개발 보다는 하나의 시스템 안에서 이들을 통합하기 위해 주로 개별 단위 분석 프로그램과 내부 및 외부 데이터베이스와의 연동, 그리고 복잡한 분석을 위해 이들 단위 분석 프로그램들 간의 연동 방식에 초점을 두고 설계하였다. 내부의 로컬 데이터베이스로부터 분석용 서열 데이터를 검색하기 위해서는 등록번호, 등록 사용자, 등록 날짜, 생물 및 유전자 이름 등 다양한 방법으로 검색이 가능하도록 설계하였고, 검색된 서열은 바로 선택된 분석 프로그램에 제공되어 분석될 수 있도록 하였다. 현재 유사 서열 검색과 문헌 검색, 연관 유전자 검색 등을 위해서는 내부의 로컬 데이터베이스 외에, GenBank, EMBL 등의 대규모 외부 서열 데이터베이스들과 Medline과 같은 외부 문헌 데이터베이스에 대한 검색 기능을 제공하도록 설계하였다. 특히 유사 서열 검색을 위해서는 검색속도가 빠른 BLAST 프로그램을 이용하였으며, 사용자의 이해도를 높일 수 있도록 BLAST의 검색결과를 재구성하도록 설계하였다. 다중 정렬을 위해서는 현재 가장 많이 이용되고 있을 뿐만 아니라 다른 후속 분석 작업과의 연동이 용이한 ClustalW 프로그램를 이용하였다. 다중 정렬을 위한 입력 서열들 역시 로컬 데이터베이스나 웹 인터페이스로부터 입력받을 수 있도록 설계하였다. 유전자 탐색을 위한 모듈로서 GenScan을 사용하였다. 유전자 기능 연구를 지원하기 위한 관련 문헌 검색기능은 PubMed시스템을 이용하도록 설계하였다. 또한 PubGene에 대한 질의가 가능하도록 구성하여 본 시스템에서 유사서열들 간의 네트워크를 생성하여 유전자의 기능을 추정할 수 있도록 설계 하였다.

GWB시스템에서는 표준화된 데이터의 상호교환을 위해 로컬 데이터베이스에 저장하고 있는 일부 또는 모든 서열 데이터들에 대해 XML 파일로 변환하는 기능을 제공하도록 설계하였다. 현재 지원하고자 하는 XML형식은 BSML (Bioinformatics Sequence Markup Language)[13]과 GAME(Genome Annotation Markup Elements)[14] 등 두 가지 형식이다. Lab Book사에 의해 개발된 BSML은 특히 DNA, RNA, 단백질 서열, 그리고 그들의 그래픽 성질들을 XML로 표현하기 위한 범용의 형식이며, 버클리 대학의 Drosophilia Genome Project 회원들과 Celera사 사이에 데이터 교환목적으로 처음 개발된 GAME은 현재 생물 서열의 특성들을 표현하기 위한 하나의 공통언어로 자리잡아 가고 있다.

3.2 서열 데이터 형식

GWB시스템은 한 실험실에서 발생하는 로컬 서열 데이터의 저장 관리 기능외에 대규모 외부

유전자 서열 데이터베이스에 대한 접근과 분석기능을 제공해야 하므로 내부적으로 다루게 될 서열 데이터 형식의 설계에 특히 주의를 기울여야 한다. 기존의 대규모 공공 서열 데이터베이스인 GenBank, EMBL, SWISPROT 등은 서로 다른 형식으로 서열 데이터를 저장 관리하고 있다. 이러한 형식들은 저마다 나름대로의 장·단점을 갖고 있으며 필요에 따라 다른 형식으로 변환되어 사용될 수 있어야 한다. 한편 대부분 소규모의 유전자 실험실에서는 실질적으로 이들 대규모 공공 서열 데이터베이스들에 비해 훨씬 간단한 형식으로 서열 데이터를 저장 관리하고 있다. 따라서 서로 다른 서열 데이터 형식들 간의 호환성을 제공할 수 있도록 서열 데이터 형식을 정하는 것은 서열 데이터 관리시스템의 설계에 매우 중요한 문제라고 할 수 있다. GWB시스템에서는 실험실 단위에서 사용하기 적합하면서도 대규모 서열 데이터베이스들과의 호환성을 고려한 새로운 서열 데이터 형식인 KSF를 설계하였다. KSF는 우선 NCBI의 BankIt에서 요구하는 형식을 수용하여 일반제출(General Submission)/참고문헌(Reference)/출처(Source) 정보, DNA 서열입력, 추가 정보를 포함하되 중규모 혹은 소규모의 유전자 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.

표 1은 KSF의 서열 데이터 형식을 표로 요약한 것이다. 대부분의 항목은 하나의 입력 값을 갖지만, 특성(Feature)와 참고문헌(Reference) 항목은 여러 개의 값을 가질 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 참고문헌(Reference)도 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 데이터 형식이외에 데이터베이스를 설계할 때와 사용자 인터페이스를 설계할 때에도 영향을 주게 된다.

〈표 1〉 KSF 데이터 형식

항 목	설 명
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 구성

GWB 시스템에서는 시스템의 특성상 두 가지 형태의 데이터베이스로 구성된다. 하나는 일반적인 사용자 및 서열 데이터의 관리와 분석에 사용되는 관계형 데이터베이스(Relational Database)이고, 다른 하나는 유사 서열 검색에 사용되는 색인 순차 파일이다. 유사 서열 검색은 DNA 서열과 단백질 서열 데이터의 분석과 관리에 필수적인 기능의 하나이다. 현재 유사 서열 검색을 위한 최고의 기술수준인 BLAST 알고리즘과 프로그램이 요구하는 데이터베이스 형식은 다름아닌 색인 순차 파일(Indexed Sequential File) 형태이다. 따라서 내부의 로컬 데이터베이스뿐만 아니라 외부의 공공 서열 데이터베이스에 대한 유사 서열 검색 기능을

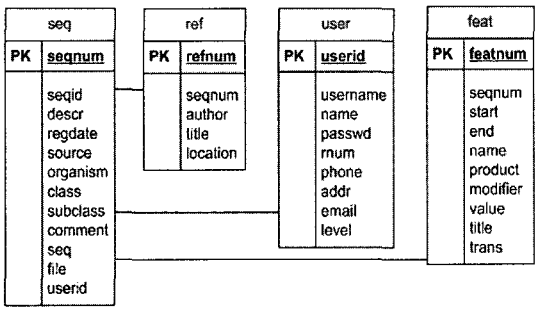
제공하고자 하는 GWB시스템에서는 이와 같은 서로 다른 두 가지 형태의 데이터베이스를 함께 운영 관리하여야만 한다.

우선 관계형 데이터베이스를 설계하기 위해서는 앞에서 언급한 KSF의 특징을 이해해야 한다. 우선 데이터의 항목을 살펴보면 크게 입력 값의 수가 정해져 있는 부분과 입력 값의 수가 정해지지 않은 부분으로 나눌 수 있다. 하나의 서열 데이터에는 기본적으로 서열 그 자체와 그 이외의 부가(Annotation) 정보가 있으며, 또한 이 부가 정보는 입력 값의 수가 일정하지 않다. 그렇기 때문에 순수 서열 정보를 제외한 나머지 참고문헌(Reference)/사용자(User)/특성(Feature) 정보는 별도의 테이블로 관리하여야 한다. 하나의 서열 데이터에는 여러 개의 참고문헌(Reference)들이 존재할 수 있고, 또한 한명의 사용자는 여러 개의 서열 데이터들을 소유할 수 있으며 이들에 대한 접근이 가능하다. 그림 7은 관계형 데이터베이스의 간단한 스키마를 나타내고 있다.

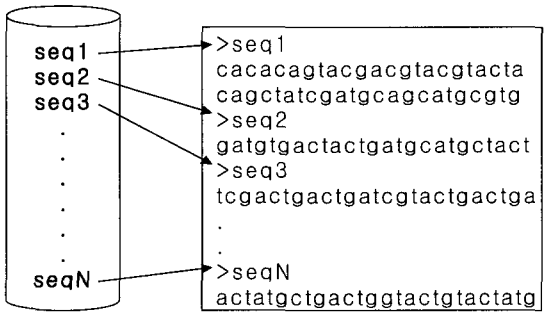
GWB시스템에는 또 하나의 데이터베이스인 색인순차 파일이 존재하는데 이것은 FASTA형식의 서열 텍스트 파일과 색인으로 구성된다. 그림 8은 BLAST를 이용한 유사서열 검색에 사용되는 색인 순차파일을 나타내고 있다. 이것은 유사 서열 검색 프로그램인 BLAST가 요구하는 고유의 데이터 형태이므로, BLAST의 실행을 위해서는 반드시 로컬 서열 데이터들을 이와 같은 형태의 색인순차파일로 저장하고 있어야 한다. 따라서 다양한 부가정보를 포함하는 각 유전자 서열 데이터에 대해, 가변길이의 문자열(Character String)로 표현되는 각 유전자 서열 자체는 BLAST에서 요구하는 색인 순차파일의 한 레코드로, 나머지 부가정보들은 모두 관계형 데이터베이스내의 튜플(Tuple)들로 나뉘어 저장 관리 되어야 한다. 즉 로컬 서열 데이터베이스는 색인 순차파일과 관계형 데이터베이스로 나뉘어 구성되고, 서열 데이터의 신규 삽입과 삭제, 갱신작업들이 필요할 때는 공유 정보인 서열 식별자(Sequence Identifier)를 이용해 관계형 데이터베이스와 색인 순차파일 양쪽 모두에서 관련 부분을 변경할 수 있도록 설계하였다. 이렇게 함으로써 두 가지 데이터베이스를 운영할 때 발생할 수 있는 데이터의 불일치성을 방지할 수 있다.

3.4 기존 프로그램들과의 연동

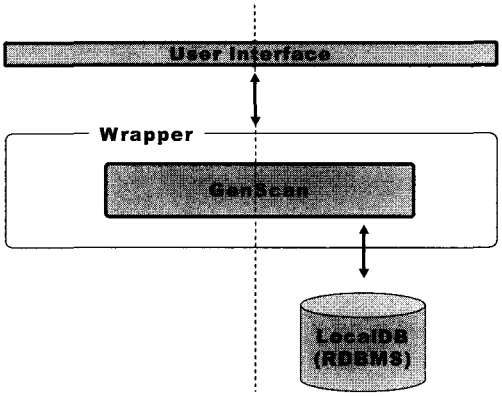
GWB시스템에서 제공하는 서열 데이터에 대한 다양한 분석(Sequence Analysis) 기능을 위해 각기 독립된 분석 프로그램들과 데이터베이스, 그리고 웹 인터페이스를 연동하는 방식을 설계하여야 한다. 대부분의 기존 서열분석 프로그램들은 단일 기능만을 제공하는 효율성이 높은 단위 프로그램들로서, 각기 서로 다른 프로그래밍 언어로 구현되어 있고, 입출력 데이터 형식이 다르다. 따라서 이들을 유기적으로 연동 가능한 하나



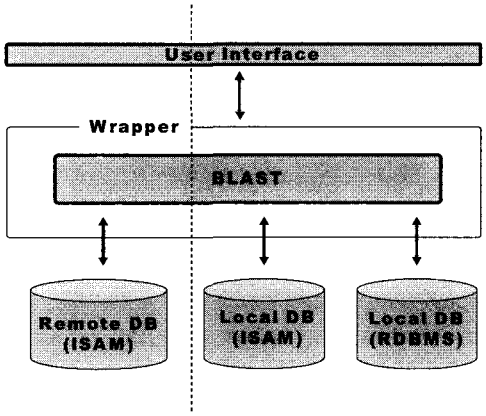
〈그림 7〉 관계형 데이터베이스의 스키마



〈그림 8〉 서열 데이터의 색인 순차 파일



〈그림 9〉 GenScan과 연동



〈그림 10〉 BLAST와 연동

의 시스템으로 통합하고 운영하기 위해 본 시스템에서는 각 분석 프로그램에 대한 포장기(Wrapper)를 설계하고 구현하였다.

그림 9는 유전자 탐색에 널리 이용되고 있는 기존의 단위 분석 프로그램인 GenScan을 연동을 하기 위한 구조를 나타내고 있다. GWB시스템에서는 사용자 인터페이스나 로컬 데이터베이스로부터 GenScan에 대한 입력 서열 데이터를 제공하고, 그 분석결과를 다시 로컬 데이터베이스나 웹 기반의 사용자 인터페이스를 통해 사용자에게 제공해주는 공통된 인터페이스 부분이 필요하다. 이와 같은 기능을 하는 것이 포장기(Wrapper)이다. 포장기는 Perl 프로그래밍 언어로 구현하며, 쉘(shell) 상태에서 실행되는 GenScan에게 쉘 상태에 맞는 형태로 입력 데이터를 전달하고, GenScan의 결과를 받아들여 웹에 적합한 형태로 결과를 재구성하여 사용자 인터페이스에 전달한다. 그림 9에서 가운데 점선을 기준으로 좌측과 우측은 각각 외부 사용자와 내부 사용자의 경우를 나누어 GenScan에 대한 입력의 차이를 표현하고 있다. 로컬 데이터베이스에 대한 접근권한이 없는 외부 사용자들은 GenScan에 필요한 입력 서열들을 웹 인터페이스를 통해 사용자가 직접 입력할 수 있으나, 반면에 내부 사용자들은 이와 같은 직접 입력방식외에 검색을 통해 로컬 데이터베이스내의 특정 서열을 찾아 GenScan의 입력 서열로 제공할 수 있다.

그림 10은 유사 서열 검색에 사용되는 BLAST 프로그램을 연동하기 위한 구조를 나타내고 있다. BLAST 프로그램을 위한 포장기(Wrapper)는 GenScan의 경우와 마찬가지로 BLAST를 위한 입출력 인터페이스를 제공한다. 즉, 웹 기반의 사용자 인터페이스를 통해 받아들인 하나의 질의 서열(Query Sequence)을 BLAST프로그램의 입력으로 제공하고, BLAST로 하여금 로컬 데이터베이스나 외부의 공공 서열 데이터베이스로부터 질의 서열과 유사한 서열 데이터들을 찾도록 한다. 그리고 BLAST의 검색결과를 받아 다시 사용자 인터페이스를 통해 그 결과를 출력해주는 역할을 수행한다. 이때 포장기(Wrapper)는 BLAST의 검색결과로서, 유사 서열들뿐만 아니라 질의 서열과의 정렬(alignment)결과도 그래픽화하여 사용자에게 보여주어야 한다. 이때 외부 사용자들은 외부의 공공 서열 데이터베이스에 대한 BLAST 검색만을 제공하나, 권한이 있는 내부 사용자들은 관계형 데이터베이스와 색인 순차파일로 구성된 로컬 서열 데이터베이스에 대한 BLAST 검색도 추가적으로 제공한다. 그림의 가운데 점선은 이와 같이 사용자별로 BLAST 프로그램이 접근 가능한 데이터베이스의 종류를 나타내고

있다.

다중 서열 정렬(Multiple Alignment)을 위해서는 역시 그림 11과 같은 ClustalW 프로그램과의 연동구조가 필요하다. 이 경우 포장기(Wrapper)는 ClustalW 프로그램에 대한 입력과 출력을 제공하며, 다중 정렬을 원하는 입력 서열 데이터들은 사용자가 웹 인터페이스를 통하여 직접 입력하거나 로컬 데이터베이스 검색을 통하여 입력할 수 있다. ClustalW가 내놓는 다중 서열 정렬의 결과는 다시 포장기(Wrapper)에 의해 웹기반의 사용자 인터페이스로 출력된다. 그림 11에서 ClustalW를 감싸고 있는 것이 포장기 부분이며, 가운데 점선은 외부 사용자 및 내부 사용자의 접근권한에 따른 구분을 나타낸다.

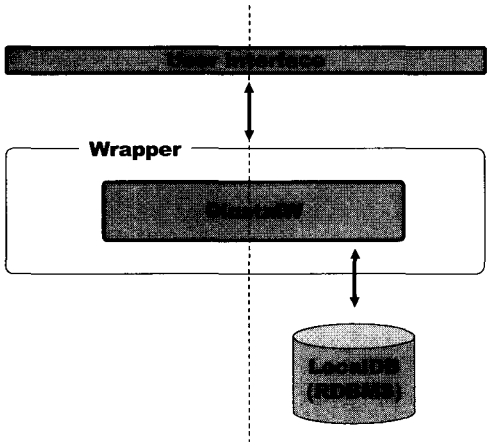
4. 시스템 구현

4.1 구현 환경

GWB시스템의 하드웨어 및 소프트웨어 개발 환경을 요약하면 표 2와 같다. GWB 시스템은 Linux 상에서 관계형 데이터베이스 관리시스템인 MySQL과 서열 데이터 와 같은 문자열 처리에 뛰어난 프로그래밍 언어인 Perl을 이용하여 구현하였다. 웹 기반의 사용자 인터페이스를 제공하기 위해 웹 서버인 Apache를 사용하였고, 서열 분석을 위해 GenScan, BLAST, ClustalW 등의 다양한 독립 분석 프로그램들과 BioPerl, BioXML 등의 Perl 라이브러리 패키지들을 사용하였다. 또 사용자의 이해를 돕고 편의성을 높이기 위해 그래픽 라이브러리 GD를 이용하여 가능한 사용자 인터페이스의 많은 부분을 그래픽화 하였다.

4.2 서열 데이터의 관리부

본 절에서는 구현된 GWB시스템의 서열 데이터 관리부의 기능과 세부사항에 대해 살펴본다. GWB시스템에서는 앞서 설계한 바와 같이 사용자별로 이용 가능한 기능을 달리 제한하고 있다. 이를 위해 로인 인증과정을 거친 내부 사용자와 그렇지 않은 외부 사용자에게는 서로 다른 주 메뉴(main menu)들이 제시되도록 구현되었다. 그리고 권한이 있는 내부 사용자에 한해 KSF 형식에 맞추어 서열 데이터를 입력하고 관리할 수 있도록 시스템을 구현하였다. 서열 데이터의 입력, 저장, 검색, 갱신, 삭제와 같은 대부분의 서열 데이터의 관리 기능은 관계형 데이터베이스인 MySQL

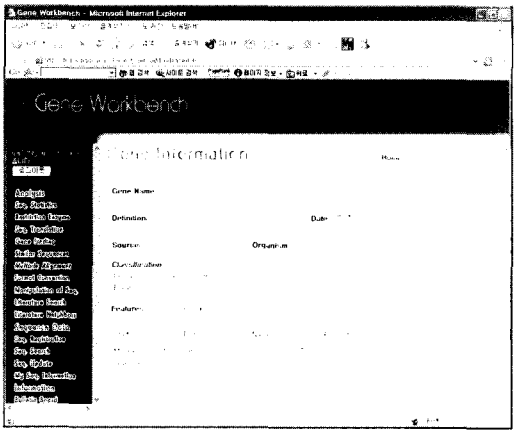


〈그림 11〉 ClustalW와 연동

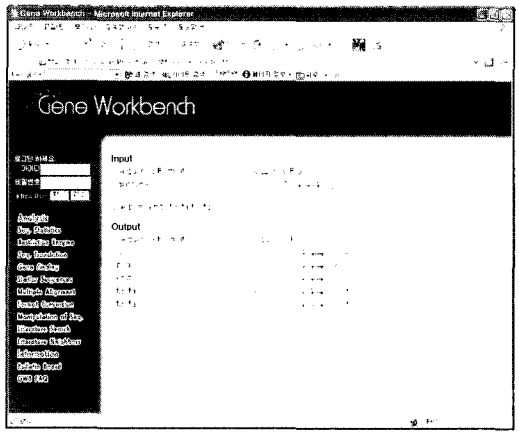
〈표 2〉 구현환경

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

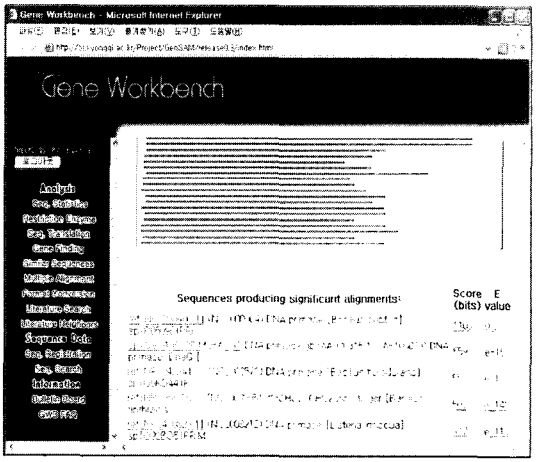
S/W 환경
- OS : Linux(Hancom 2.2)
- Language : Perl 5.6.0
- Library : Bioperl 1.0.2, BioXML, GD, Javascript
- External Package : GenSCAN, BLAST,ClustalW
- DBMS : MySQL 3.23.47
- WebServer : Apache 1.3.22



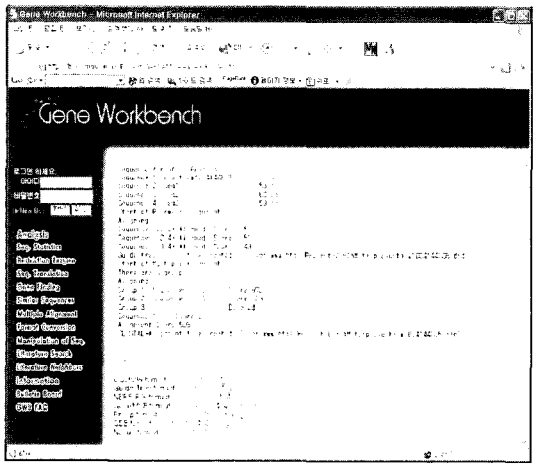
〈그림 12〉 서열 데이터의 입력



〈그림 13〉 서열 데이터의 형식 변환



〈그림 15〉 유사 서열 검색

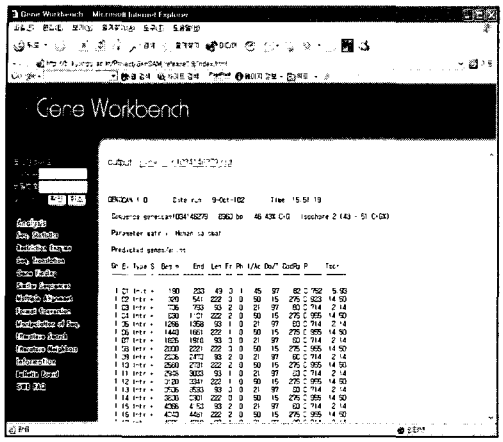


〈그림 16〉 다중 서열 정렬

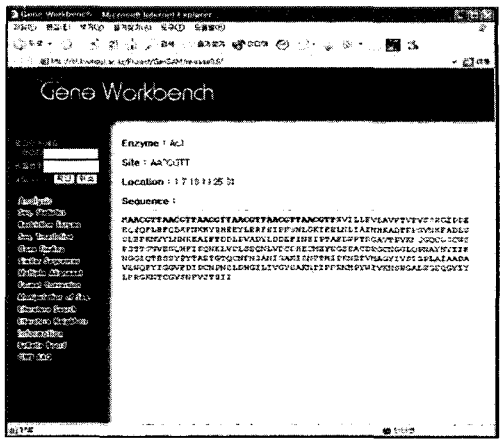
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE bx-game:game PUBLIC "game" "http://www.bioxml.org/dtds/current/game.dtd">

<bx-game:game xmlns:bx-game="http://www.bioxml.org/dtds/current/game.dtd">
    <bx-game:flavor>chunkable.</bx-game:flavor>
    <bx-seq:seq bx-seq:id="AI129902" bx-seq:length="317" bx-seq:type="dna" xmlns:bx-seq="http://www.bioxml.org/dtds/current/seq.dtd">
        <bx-seq:residues>CTCCGCGCCAACTCCCCCCACCCCCCCCCACACCCC</bx-seq:residues>
    </bx-seq:seq>
    <bx-feature:feature bx-feature:id="1" bx-feature:parent="" xmlns:bx-feature="http://www.bioxml.org/dtds/current/feature.dtd">
        <bx-feature:type>source</bx-feature:type>
        <bx-feature:score bx-feature:type="clone_lib">Soares_pregnant_uterus_NbHPU</bx-feature:score>
        <bx-feature:score bx-feature:type="dev_stage">adult</bx-feature:score>
        <bx-feature:score bx-feature:type="sex" female</bx-feature:score>
        <bx-feature:score bx-feature:type="organism">Homo sapiens</bx-feature:score>
        <bx-feature:score bx-feature:type="clone">IMAGE:1713149</bx-feature:score>
        <bx-feature:score bx-feature:type="note">Organ: uterus; Vector: pT7T3-Pac; Site_1: Not I; Site_2: Eco RI; 1st strand cDNA was
        <bx-feature:score bx-feature:type="lab_host">DH10B</bx-feature:score>
        <bx-feature:score bx-feature:type="db_xref">taxon:9606</bx-feature:score>
        <bx-feature:seq_relationship bx-feature:seq="AI129902" bx-feature:type="query">
            <bx-feature:span>
                <bx-feature:start>1</bx-feature:start>
                <bx-feature:end>317</bx-feature:end>
            </bx-feature:span>
        </bx-feature:seq_relationship>
    </bx-feature:feature>
</bx-game:game>
```

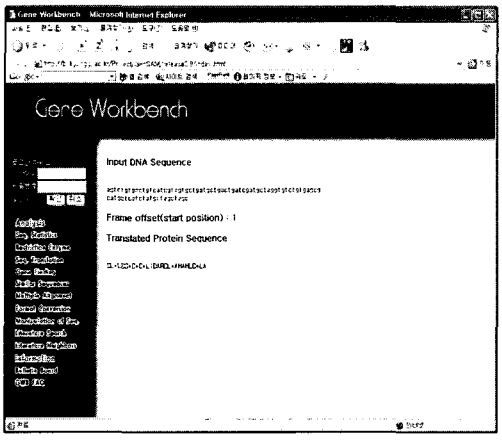
〈그림 14〉 변환된 XML 화일



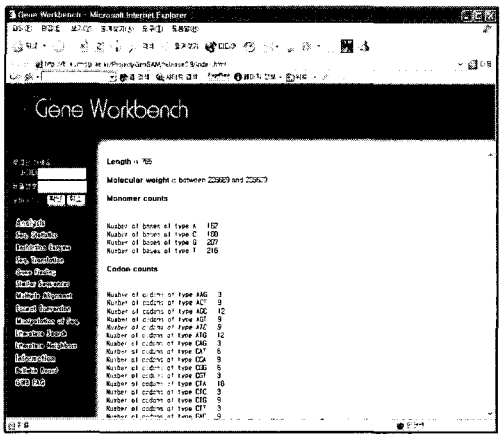
〈그림 17〉 유전자 탐색



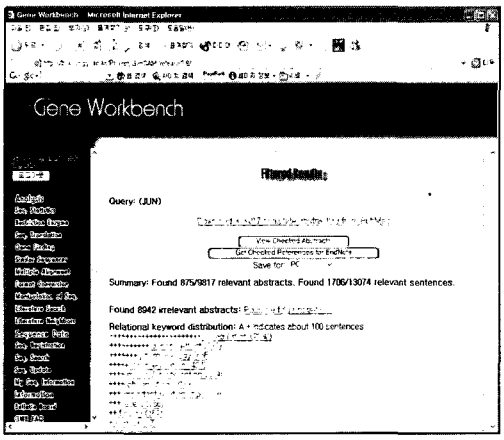
〈그림 18〉 제한 효소 탐색



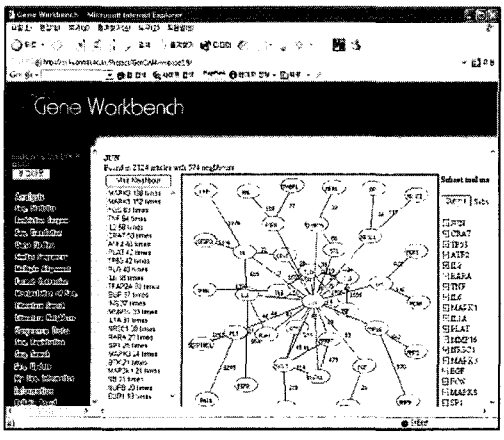
〈그림 19〉 단백질 아미노산 서열로 변환



〈그림 20〉 서열 통계 분석



〈그림 21〉 문헌 검색



〈그림 22〉 연관 유전자 탐색

과 Perl을 이용하여 구현하였다. 그림 12는 내부 사용자에게 제공되는 유전자 서열 데이터의 한 입력화면이다.

GWB에서는 주요 서열 데이터 형식간의 변환 기능을 제공하기 위해 서열 데이터 형식 변환 모듈을 구현하였다. 현재 변환 가능한 데이터 형식은 GenBank, EMBL, FASTA 등을 포함해 총 여섯 가지 형식이다. 그림 13은 하나의 서열 데이터를 서로 다른 형식의 서열 데이터들로 변환하는 화면을 보여주고 있다. 한편, GWB시스템에서는 이러한 서열 데이터 형식의 변환 기능 외에 보유 데이터의 원활한 교환을 위해 XML 파일로 변환하는 모듈을 Bioperl과 BioXML을 이용하여 추가로 구현하였다. 현재 지원되는 XML 형태는 GAME과 BSML 등 두 가지이다. 그림 14는 XML 파일로 변환된 한 서열 데이터의 예를 보여주고 있다. 또한 이렇게 XML 파일로 변환된 서열 데이터는 Genomic Viewer와 같은 뷰어를 이용하여 시각화할 수 있다.

4.3 서열 데이터의 분석부

본 절에서는 구현된 GWB시스템의 서열 데이터 분석부의 기능과 세부사항에 대해 살펴본다. 앞서 설명한 바와 같이 GWB는 유전자 서열 데이터에 대한 다양한 분석 기능을 제공할 수 있도록 설계하였으며, 몇몇 주요 분석 기능들은 기존의 우수한 단위 분석 프로그램들을 연동하여 구현하였다. 그림 15는 BLAST를 이용한 유사 서열 검색의 결과화면을 나타내고 있다. 유사 서열 검색은 질의 서열과 유사한 서열들을 찾아주는 것으로서, 질의 서열과 데이터베이스에 저장된 서열과의 유사성을 검사하여 가장 유사성이 높은 서열부터 낮은 순으로 정렬하여 사용자에 보여준다. 그림 15에서 화면 상단의 붉은색 막대그래프는 질의 서열과 대상 서열간의 유사도를 나타내는 그래프이다. 이러한 그래프 부분은 그래픽 라이브러리인 GD를 이용하여 구현하였으며, 이미지 맵 기술을 이용하여 각 그래프에 대상 서열 정보로 바로 이동할 수 있는 하이퍼링크가 연결되도록 구현하였다.

이 밖에도 GWB시스템에는 다양한 분석 기능을 구현하고 있는데, 그림 16은 ClustalW를 기초로 구현된 다중 서열 정렬 모듈의 한 실행화면을 보여주고 있으며, 그림 17은 한 서열 데이터 안에 포함된 유전자의 위치를 찾아주는 유전자 탐색 모듈의 실행화면을 보여준다. 또 그림 18은 제한 효소 탐색 모듈의 한 실행화면을, 그림 19는 한 DNA 서열을 단백질 아미노산 서열로 변환하는 실행화면을, 그림 20은 한 서열 데이터 내에 포함된 코돈의 수와 위치 등 다양한 통계치를 분석하는 실행화면을 보여준다. GWB 시스템의 중요한 특징 중 하나는 바로 유전자 기능 연구를 지원하는 기능들을 제공하는 것이다. 이를 위해 문헌 데이터베이스인 Medline에 대한 관련 문헌 검색 모듈과 문헌상 관련도가 높은 다른 유전자들을 찾아주는 PubGene시스템과의 연계모듈들을 구현하였다. 그림 21과 그림 22는 각각 문헌 검색 모듈의 한 실행화면과 연관 유전자 탐색 모듈의 한 실행화면을 보여주고 있다.

5. 결 론

본 논문에서는 유전자 서열 관리 및 분석 시스템인 GWB의 설계와 구현에 대해서 살펴보았다. GWB에서는 유전자 서열 연구에 있어서 보다 효율적인 시스템을 구현하고자 서열 데이터베이스 관리 기능과 다양한 분석 기능들을 하나의 시스템으로 통합함으로써 기존 시스템들의 단점을 보완하였다. GWB는 대규모 공공 서열 데이터베이스들과는 달리 유전자 서열 데이터를 다루는 실제 실험실 사용자의 편의성을 최대한 고려하여 설계한 시스템이다. GWB는 웹 기반의 인터페이스를 제공함으로써 기관 내부와 외부의 사용자 모두에

게 이용의 편의성을 제공하면서도, 엄격한 사용자 권한 제한 기능을 제공하여 로컬 서열 데이터베이스의 보안성을 높였다. GWB시스템 설계의 가장 중요한 이슈는 서로 상이한 분석 프로그램들을 어떻게 하나의 시스템으로 통합할 것이며, 또 이들 프로그램들이 요구하는 서로 다른 서열 데이터 및 서열 데이터베이스 형태를 제공할 수 있느냐는 것이다. GWB는 이 문제들을 해결하기 위해 공통 서열 데이터 형식인 KSF를 제안하였으며, 로컬 서열 데이터베이스를 관계형 데이터베이스부분과 색인 순차파일부분으로 나누어 구성하였으며, 서로 상이한 서열 데이터 형식간의 변환 기능과 XML 파일로의 변환 기능을 제공하도록 하였다.

계획하고 있는 향후 연구로는 다양한 실제 서열 분석작업들을 통해 GWB시스템의 효율성을 검증해보는 일과 기능 유전체(Functional Genomics) 연구를 직접 지원할 수 있는 새로운 기능들을 추가해나가는 일, 사용자에게는 보다 편의성을 제공하고 GWB시스템에게는 보다 유연한 제어를 주도록 하기 위해 GWB시스템의 각 독립적인 모듈들을 에이전트(Agent)로 변환하여 GWB시스템을 하나의 다중 에이전트시스템(Multi-Agent System)으로 재구성하는 일, 그리고 현재 여러 기관을 중심으로 활발히 진행되고 있는 데이터 표준화에 GWB시스템을 맞추어 나가는 일 등이 있다.

참 고 문 헌

[1] Altschul,S.F. W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. "A Basic Local Alignment Ssearch Tool". Journal of Molecular Biology, Vol. 215, No. 3, pp. 403-410, Oct 1990,

[2] Altschul, S.F., et. al., "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs". Nucleic Acids Res. Vol.25, pp. 3389-3402, 1997

[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", Proceedings of the German Conference on Bioinformatics(GCB-96), Leipzig, Germany, pp. 44-51, 1996

[4] Bairoch A., Apweiler R, "The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL", pp. 21-25, Oxford University Press, 1996

[5] Burge, C. B. and Karlin, S. "Finding the Genes in Genomic DNA". Current Opinion in Structural Biology, Vol.8, pp. 346-354, 1998

[6] Cynthia Gibas, Per Jambeck, "Developing Bioinformatics Computer Skills", O'Reilly, 2001

[7] David W. Mount, "Bioinformatics : Sequence and Genome Analysis", CSHL, 2001

[8] Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L., "Fast algorithms for Large-Scale Genome Alignment and Comparison", Nucleic Acids Research, Vol.30, No.11 pp. 2478-2483, 2002

[9] Des Higgins, "Bioinformatics: Sequence, Structure and Databanks", Oxford University Press, 2000

[10] Gotoh, O., "Alignment of Three Biological Sequences with an Efficient Traceback Procedure", Journal of Theoretical Biology, Vol.121, pp. 327-337, 1986

[11] 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, Vol.7, No.11, pp. 957-959, 2000

[12] Henikoff, S., and J.G. Henikoff, "Amino Acid Substitution Matrices from Protein Blocks". Proceedings of the National Academy of Sciences of the USA, Vol.89, No.22, 10915-10919, 1992

[13] Lab Book Inc., "BSML Reference Manual", 2002.

[14] Lewis, S.I. et al, "AN Intelligent System for Interpretation of Sequence Alignment Results", The Institute of Genomic Research: Second Annual Conference on Computational Genomics, 1998.

[15] Morgenstern B., Dress A. and Wener T., "Multiple DNA and Protein Sequence Alignment Based on Segment-To-Segment Comparison", Proc. Natl. Acad. Sci. USA, Vol.93, pp. 12098-12103, 1996.

[16] Needleman, S.B. and Wunsch C.D., "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", Journal of Molecular Biology, Vol. 48, pp.443-453, 1970.

[17] Smith T.F. and Waterman M.S., "Comparison of Biosequences", Advanced Applied Mathematics, Vol.2, pp.483-489, 1981.

[18] Taylor W.R., "A Flexible Method to Align Large Numbers of Biological Sequences", J. Mol. Evol., Vol.28, pp. 161-169, 1988.

[19] Thompson J., Higgins D. and Gibson T., "CLUSTAL W: Improving the Sensivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice", Nucl. Acids Res., Vol.22, pp.4673-4690, 1994.

[20] Wang L. and Jiang T., "On the Complexity of Multiple Sequence Alignment", J. Comput. Biol., Vol.1, pp.337-348, 1994.

◐ 저 자 소 개 ◑

김 인 철

1985년 서울대학교 수학과 졸업(학사)
1987년 서울대학교 대학원 전산과학 졸업(석사)
1995년 서울대학교 대학원 전산과학 졸업(박사)
1996년~현재 경기대학교학 정보과학부 조교수, 부교수
관심분야 : 인공지능, 데이터마이닝, 바이오인포매틱스
E-mail : kic@kyonggi.ac.kr

진 훈

1998년 경기대학교 전자계산학과 졸업(학사)
2000년 경기대학교 대학원 전자계산학과 졸업(석사)
2000년~현재 경기대학교 전자계산학과 박사과정 재학 중
관심분야 : 인공지능, 데이터마이닝, 바이오인포매틱스
E-mail : jinun@kyonggi.ac.kr