웹 기반의 유전자 서열 분석 및 관리 시스템

허진석[0], 김현식, 예형석, 진훈, 김인철
경기대학교 전자계산학과
{hjs0823[0], advance7, elta, jinun, kic}@kyonggi.ac.kr

Gene Sequence Analysis and Management System based on web

Jin-Seok Heo[0], Hyun-Sik Kim, Hyung-Seok Ye, Hoon Jin, In-Cheol Kim
Dept. of Computer Science, Kyonggi University

요 약

본 논문에서는 하나의 시스템안에서 효율적인 유전자 데이터의 관리와 다양한 서열 분석작업이 가능한 웹 기반의 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.

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 대한 관심도 따라서 높아졌다. 또한 인간을 비롯해 많은 생명체의 전체 또는 일부의 유전자 염기서열이 밝혀진 현재, 이러한 유전자 염기서열을 바탕으로 각 '유전자의 구조와 기능을 밝히는 구조/기능 유전체학(Structural/Functional Genomics)과 개체간의 차이, 환경의 대한 차이 등을 연구하는 비교 유전체학(Comparative Genomics)에 관한 연구도 활발히 진행되고 있다. 따라서 생물정보학 분야도 한 생명체의 새로운 유전자를 찾거나 유전자 서열 데이터 관리하는 수준에서 벗어나 이러한 구조 및 기능 유전체학을 지원할 수 있는 정보시스템을 구축하는 연구에 관심을 쏟기 시작했다.

본 논문에서는 유전자 데이터를 다루는 일반적인 생명과학실험실에서 자체적으로 유전자 서열 데이터를 저장 관리하면서, 다양한 서열 분석작업을 수행해볼 수 있는 소프트웨어시스템을 설계 구현하고자 한다. 오랫동안 기존의 생명과학 실험실에서의 데이터 관리는 정확한 표준안이 마련되어 있지 않은 상태에서 일정한 양식을 정하여 그 양식에 맞도록 작성된 텍스트 파일형태로 이루어져왔다. 또 서열분석에 사용되는 분석프로그램의 종류도 매우 다양하며, 처리 데이터 양식이나 수행환경, 구현 프로그래밍 언어도 조금씩 차이가 있고 프로그램간의 연동도 되지 않아 분석작업이 어렵고 불편하였다. 본 논문에서는 이와 같은 문제점을 보완하여 하나의 시스템안에서 효율적인 유전자 데이터의 관리와 다양한 서열 분석작업이 가능한 웹 기반의 서열 분석 및 관리 시스템인 GWB(Gene Workbench)를 설계하고 구현하였다. GWB는 로컬데이터베이스 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 가지고 있다.

2 유전자 서열 분석

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 염기서열에서 intron을 제거한 exon만으로 구성된 mRNA가 생성된다. 이 때 T는 U(우라실)로 치환된다. 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. 아래의 (그림 1)은 염기 서열에서 mRNA 생성하여 ORF를 찾아내는 과정이다. 이와 관련된 도구로는 GENSCAN, GRAIL등이 있다.

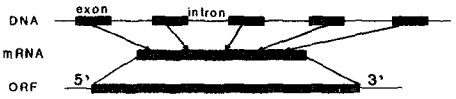

그림 1 유전자 탐색

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. 아래의 (그림 2)은 상동성 검

색을 위한 서열 짝 정렬을 나타내고 있다.

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

그림 2 유사 서열 검색

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

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

그림 3 다중 서열 정렬

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. 예를 들어 아래의 (그림 4)는 GenBank와 EMBL의 데이터인데 서로 다른 형식임을 알 수 있고, 나타내고 있는 정보중 많은 속성들이 소규모 실험실에 적합한 형태가 아님을 알 수 있다. 그리고 또한 아직까지도 서열 관련 데이터베이스에 관한 표준안이 없기 때문에 서로 다른 형식의 데이터베이스들이 존재한다.

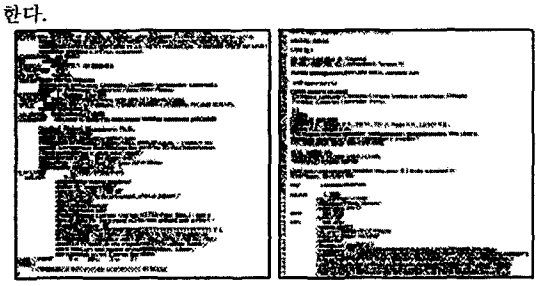

그림 4 데이터 형식 : GenBank(좌), EMBL(우)

3. 설계

3.1 서열 데이터 형식

생물학에서 다루어지는 서열들은 (그림 4)에서 보듯이 GenBank, SWISSPROT, EMBL 등 보유 데이터베이스마다 여러 가지 형식으로 존재할 때가 많다. 이러한 형식들은 저마다 나름대로의 장·단점을 갖고 있으며 필요에 따라 다른 형식으로 변환되어 사용되어야 한다. 또한 실질적으로 소규모의 실험실에서는 서열에 대한 정보를 위와는 다른 형식으로 훨씬 간단하게 저장하여 관리하고 사용한다. 그러므로 데이터 형식들 간의 호환성을 제공하도록 하는 것은 생물학 정보를 시스템을 통해 관리하는데 있어서 필수불가결한 기능이라 할 수 있다. GWB에서는 사용하기 적합하면서도 유명 서열 형식과의 호환성을 고려한 새로운 서열정보 형식(KSF, Kyonggi Sequence Format)을 고안하기로 하였다. KSF는 우선 NCBI의 BankIt에서 요구하는 형식을 수용하여 General Submission /Reference/Source 정보, DNA 서열입력, Additional 정보를 포함하되 소규모의 실험실에 적합하도록 최대한 간략화시켰다. 이를 우리가 대상으로 하였던 유전학 실험실에서 기록되어 오던 기존의 서열정보 형식과 통합시킴으로써 수기로 작성됨으로 인해 손실이 쉽고 부실하게 기록되어 오던 서열정보를 재대로 관리할 수 있도록 하였다. KSF를 사용할 경우 기존의 유명 서열 형식들과 호환가능할 뿐만 아니라 표준화 연구가 진행중인 BSML(XML)형태로의 변환기능을 제공한다.

3.2 데이터베이스 설계

GWB 시스템에는 크게보면 3가지의 데이터베이스로 구분할 수 있다. 하나는 서열 및 사용자 데이터베이스, BLAST용 데이터베이스, 외부 데이터베이스로 나눌수 있다. 서열 및 사용자 관리에 사용되는 데이터베이스는 모두 4개의 테이블로 구성되어 있으며, 이중 두개는 등록대기 서열과 등록 대기 사용자를 위한 용도이고, 나머지 두개는 관리자가 등록을 승인하면 나머지 2개의 테이블로 옮겨지게 된다. 이것은 내부 사용자라 할지라도 데이터를 임의로 수정하거나 삭제할 경우 데이터의 질이 낮아질수 있기때문이다. 서열정보를 크게 필수정보(서열의 일반정보, DNA서열, 소스정보, 입력자 정보)와 부가정보(논문 정보, 서열 특성, 주석)으로 분리하여 데이터베이스의 각 필드별로 KSF 화면을 통해 입력된 서열 정보를 저장하였다.

RID	UID	SEQ	ANNOT	GeneID	Rdate	Author	Subject	JName	Source	GWB

표 1 Gene_Info 테이블

(표 1)에서 GRID는 자동으로 증가되는 서열의 고유한 등록번호이고, UID는 사용자 계정, SEQ는 실제 서열부분이며, GWB는 서열 데이터 파일을 저장하는 필드이다.

GUID	UID	Passwd	RNumber	HPhone	Email	Address	URDate

표 2 User_Info 테이블

(표 2)는 사용자 테이블로 UID를 통하여 사용자에 따른 권한을 조정한다. 유사서열 검색 작업을 수행하기 위해서는 별도로 BLAST에 적합한 자료형태를 제공하여 한다. 이때 인덱싱 프로그램인 formatdb를 사용하여 입력을 FASTA 형식의 파일을 받아서 BLAST에 적합한 형태의 인덱스 파일들(*.nhr, *.nin, *.nsd, *.nsq, *.nsi)을 생성한다. GWB에서는 입력받은 서열정보를 가지고 다음과 같이 3가지의 타입의 결과 파일을 생성토록 하였다.

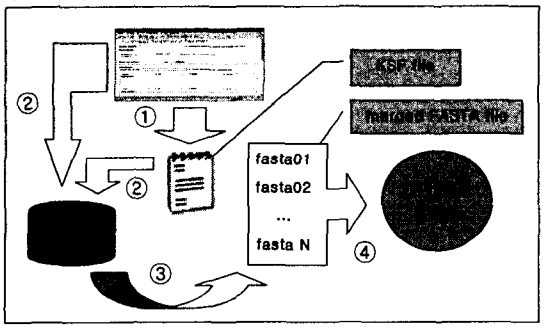

그림 5 서열등록 과정

(그림 5)에서 입력된 서열은 ①과정을 통해 먼저 KSF 서열형식 정보파일을 생성하고 ②과정을 통해서 관계형 데이터베이스에 저장된 후 ③과정을 통해 머지(merge)된 FASTA파일을 생성하고 최종으로 BLAST 프로그램 수행을 위한 색인파일들이 생성된다.

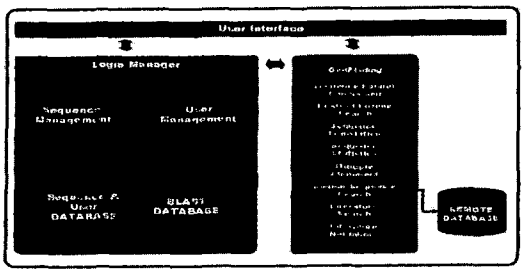

그림 6 시스템 구조

3.3 시스템 구성

시스템은 크게 서열 관리, 사용자 관리, 유전자 서열 분석으로 구성되어 있으며 구조도는 (그림 6)와 같다. 서열 및 사용자 관리는 Login Manager를 통해서만 접근할 수 있으며, Login 후에는 쿠키를 사용하여 관리한다. 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. 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. 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 LIMS내에서 유사서열들 간의 네트워크를 생성하여 유전자의 기능을 추정할 수 있도록 하였다. 또한 각 데이터베이스의 형식에 맞도록 데이터베이스의 형식을 변환하는 기능을 가지고 있다.

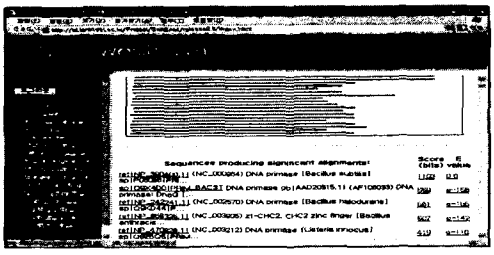

그림 7 유사서열 검색

4. 구현

GWB는 사용자의 작업능력을 높이며 위치에 구애받지 않고 이용할 수 있도록 웹 기반으로 개발되었다. 시스템 구현환경은 아래와 같다.

■H/W : 1.8GHz Dual CPU, 1Ghz Memory, 50Gbyte HDD
■S/W : Linux 7.2, Perl 5.8, Bioperl 1.02

GWB의 모든 기능을 사용하기 위해서는 반드시 로그인한 후에 사용하여야 한다. (그림 7)은 내부 사용자의 유사서열 검색전 장면이다. 그림 상단에 붉은 선은 질의 서열과 검색된 서열의 유사도를 그래프의 형태로 나타낸 것이다. 붉은 선 하나하나는 하나의 유전자를 나타내며, 이 부분에 링크를 붙여 서열로 이동할 수 있도록 구현하였다. 내부 사용자는 유사서열 검색시에 Local 데이터베이스와 Remote 데이터베이스 모두를 사용할수 있지만, 외부 사용자는 Remote 데이터베이스 만을 사용할 수 있다. 이것은 내부의 데이터를 보호하기 위해서 내부 사용자에게만 접근을 허용하는 것이다. 서열을 등록하기 위해서는 서열등록 형식에 맞춰 작성한 후 서열을 등록을 요청한다. 그후 관리자가 검토후에 등록을 하게 된다. (그림8)은 등록된 서열을 검색하여 삭제하려는 화면이다. 이 기능은 관리자에게 부여된 권한이다.

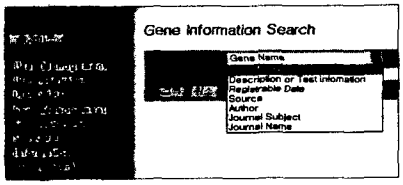

그림 8 관리자 모드

(그림 9)는 문헌 정보를 이용하여 유사한 유전자들의 관계를 네트워크로 나타내고 있다.

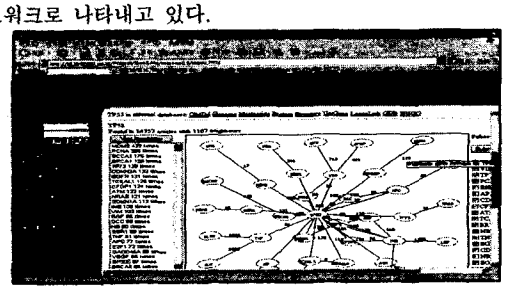

그림 9 문헌정보를 이용한 관련 유전자 탐색

5. 결론

본 논문에서는 생명과학 실험실에서 발생하는 유전자 데이터를 하나의 시스템안에서 효율적으로 관리하고 다양한 서열 분석작업을 수행할 수 있는 웹 기반의 서열 분석 및 관리 시스템인 GWB를 설계하고 구현하였다. 향후계획으로는 DNA서열 수준의 정보관리 뿐만 아니라 단백질 형태의 정보를 관리하고 이용할 수 있는 시스템으로의 업그레이드를 시도하는 것이다.

참고문헌

[1] Cynthia Gibas, Per Jambeck, Developing Bioinformat-ics Computer Skills, O'Reilly, 2001
[2] David W. Mount, Bioinformatics : Sequence and Ge-nome Analysis, CSHL Press, 2001.
[3] Des Higgins, Bioinformatics: Sequence, Structure and Databanks, Oxford University Press, 2000.