유전자 데이터베이스의 설계 및 구현: streptomyces data를 예로

김진[0] 김범준[1] 김정미[2] 김동회[0]
한림대학교 정보통신공학부[0]
제주대학교 의과대학 미생물학교실[1]
(주)바이오메드랩 부설연구소 마이크로어레이센터[2]
jinkim@sun.hallym.ac.kr, kbumjoon@cheju.cheju.ac.kr, jkjeong@bio.bmelab.com,
dhkim@sun.hallym.ac.kr

Design and Implementation of gene sequence database with streptomyces data

Jin Kim[0] Bum-Joon Kim[1] Jeongmi Kim[2] Donghoi Kim[0]
Div. of computer and engineering, Hallym Univ.[0]
Dept. of Microbiology, Cheju National Univ.[1]
Biomedlab Institute[2]

요 약

유전자의 서열 및 관련 정보가 폭발적으로 증가함에 따라, 사용자들에 대한 유전자정보 서비스, 온라인 상에서의 효율적인 서열정보 분석, 서열정보에 대한 효율적인 관리, 관련된 연구자들과의 정보공유 등이 필요하게 되었다. 본 논문에서는 인터넷 상에서 streptomyces 유전자 data를 효율적으로 관리하는 한편, 사용자들에게 유용한 서비스를 제공하는 시스템의 설계 및 구현에 관하여 논의하였다. 사용자는 본 시스템으로부터 원하는 유전자 정보를 다운로드 받을 수 있다. 또한 분석을 원하는 유전자를 streptomyces database내의 유전자들과 비교하여 유용한 정보를 추론할 수 있다.

1.서론

생물학 역사상 가장 중요한 프로젝트의 하나인 Human Genome Project(HGP)[1]로 말미암아 생명현상을 이해할 수 있는 기본적인 자료인 서열 및 유전자 정보는 가하급수적으로 증가하게 되었다. 이러한 정보가 증가함에 따라, 관련 community에 대한 서열정보 서비스, 온라인 상에서의 효율적인 서열정보 분석, 서열정보에 대한 효율적인 관리, 관련된 연구자들과의 정보공유가 절실하게 요구되고 있다. 따라서 가장 효율적으로 이와 같은 목표를 달성하기 위하여 인터넷 상에서 해당 유전자 정보 데이터베이스의 관리 및 서비스를 제공할 수 있는 웹 사이트의 개발이 필요하게 되었다. 미국을 비롯한 정보 강국에서는 이러한 유전자 정보를 효율적으로 관리하고, 서비스를 제공하는 사이트들이 만들어져 널리 사용되고 있다. 그러나 국내에서는 이러한 유전자 정보 관리 시스템이 드물어 여러 가지 문제점을 나타내고 있다.

본 논문에서는 방선균 유전자를 관리할 수 있는 최초의 국제적인 서열 데이터베이스 관리 시스템의 설계 및 제작에 관하여 논하였다.

본 논문은 다음과 같이 구성되었다. 2장에서는 서열 및 유전자 데이터베이스 시스템을 설명하였으며, 3장에서는 온라인상에서 유전자 정보를 영어로 제공하는 최초의 한국형 서열 데이터 관리 시스템인 Streptomyces Data Management System(SDMS)의 구현에 관하여 설명하였다. 마지막으로 5장에서는 본 논문의 결론 및 향후과제를 설명하였다.

2. 서열 및 유전자 데이터베이스 시스템

2.1 Human Genome Project

Human Genome Project는 미국정부와 미국국립위생연구소(National Institutes of Health; NIH)에 의해 13년 이라는 기간동안 계획하여 1990년부터 시작하였고 빠른 유전공학기술이 진보함으로써 앞으로 2003년까지 완전한 자료획득을 목표로 가속화하고 있다

HGP의 목표는 다음의 4가지로 요약될 수 있다. 첫째, 인간의 유전자 4만개의 정보확인. 둘째, 인간의 DNA를 이루고 있는 30억 개의 화학적염기배열을

결정. 셋째, 데이타베이스 정보의 기록. 넷째, 데이타분석의 기술상의 문제를 개발. 다섯째, 프로젝트에 관한 도덕적, 법률적, 사회적인 이슈에 대한 설명 등이다.

프로젝트의 목표를 성취하기 위해 연구가들은 인간이 아닌 생물들의 유전적인 연구를 하고 있다. 이러한 것은 인간의 장내에 기생하고 있는 대장균, 파일파리, 그리고 laboratory mouse를 포함하고 있다.

생물의 유전자를 포함하고 있는 DNA 모두를 게놈이라고 한다. 유전자는 모든 생물의 단백질을 생산하는데 필요한 정보를 가지고 있다. 단백질은 생체 내에서 식품의 이화, 감염에 대한 저항성, 어떤 물질의 구성, 때때로 어떻게 작용하는지 어느 정도 알려진 상태이다. DNA는 4개의 화학적으로 비슷한 염기로 이루어져 있다. 이것은 게놈 상에서 수 백만, 수 억만 번이나 반복되어 있다. 인간의 게놈을 예로 들면 3억 쌍의 염기결합을 하고 있다. 이들 4개의 특정순서는 매우 중요한 것이다. 생물은 종에 따라 각기 따른 게놈을 가지고 있다. 이러한 각기 다른 게놈을 밝히는 게 HGP의 핵심이라 할 수 있다. 모든 생물체는 DNA의 상동적인 배열을 가지고 있기 때문에 무생물로부터 인간의 생물학적 과학지식 얻을 수 있을 거라 기대된다.

2.2 유전자 데이터베이스

유전자 데이터베이스는 생명공학에 있어 필수 불가결한 요소인 유전자데이타를 관리하며, 관련 데이터를 체계적이며 지속적으로 수집 보존하도록 하며, 이들로부터 유용한 정보를 추출할 수 있도록 한다. 데이터베이스내의 데이터들은 관련정보와 함께 효율적으로 인터넷 상으로 서비스될 수 있도록 하고 있다.

2.3 Streptomyces Data

일반적으로 항생물질은 Tyndall에 의해서 미생물 상호간의 길항 작용이 관찰된 후 1929년 Fleming에 의해 발견된 penicillin을 기초로 하여 Waksman과 Woodruff가 *Streptomyces antibioticus*에서 streptomycin을 분리하여 폐결핵의 치료약으로 사용하면서 Streptomyces 속을 포함한 방선균류는 항생물질의 생산균으로서 산업적으로나 중요한 미생물로 다루어지게 되었다. 즉, streptomycin이 발견된 이후로 많은 연구자들에 의해 방선균을 항생물질 생산균주의 대상으로서 연구를 하기에 이르렀고, 그 결과 많은 항생물질이 발견되었다. 그 후 방선균으로부터 항생물질 외에 여러 가지 생리활성 물질들이 계속 분리되면서 방선균에 대한 새로운 항생물질의 탐색이 진전되고 있으며, 최근에는 재조합을 이용한 항생제 생산에 연구도 이루어지고 있는 실정이다.

방선균은 2차 대사 산물의 다양성으로 인하여 생리활성 물질의 종류도 다양하다. 지금까지 미생물로부터 탐색된 10,000여종의 생리활성 물질 가운데 약 2/3에 해당되는 64% 정도가 방선균으로부터 발견되었으며, 세균으로부터는 약 13%, 곰팡이로부터는 약 23%가 발견되어 각종 생리활성 물질의 탐색에 있어서 방선균이 차지하는 비중은 매우 크다. 이러한 이유 때문에 생물 소재 산업에 있어서 산업적으로 가장 중요한 미생물로 부각되고 있다. 이러한 방선균을 분류하는데 가장 중요한 유전자가운데 하나가 길이 342base pair로 이루어진 *rpoB*라는 유전자이다. 우리는 다양한 방선균으로부터 *rpoB* 유전자를 해독하여 데이터베이스에 저장하여 유용한 서비스를 제공하려 한다.

3. Streptomyces Data Management System(SDMS)

3.1 구현환경

방선균 연구에서 개발된 시스템은 WWW상에서 일반 사용자들이 손쉽게 정보검색 및 유전자 정보를 얻을 수 있도록 구현되었다. 모든 정보는 영어로 표현되어 있어 영어를 사용할 수 있는 사용자들이 사용할 수 있도록 한다.

사용자가 클라이언트(브라우저)를 통해 본 서버에 접속하면 common Gateway Interface(CGI)는 사용자가 클라이언트를 통해 서버로 보낸 데이터를 서버에서 작동중인 데이터처리프로그램에 전달하고, 프로그램에서 처리된 데이터를 다시 서버를 거쳐 클라이언트에게 되돌려보낸다. 본 시스템의 개발을 위해 사용된 도구는 php와 MySql이다.

3.2 기능

이 시스템의 주요구성은 News, Online analysis, Data Download, Contacts, Citation, Link, Board, Help등으로 구성되어 있다. 그림 1은 본 연구에서 개발한 SDMS의 최초화면으로 접속 URL은 www.streptomyces.net이다.

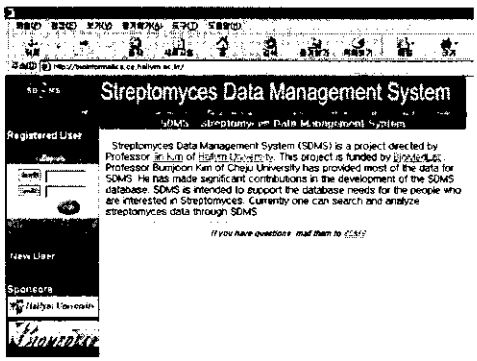

그림 1

사용자는 중요 서비스(Data download, Online analysis)를 제공받기 위하여 사용자의 정보를 등록하고 login ID, password를 부여 받아야 한다. 그림 2는 사용자의 정보를 등록하는 화면이다.

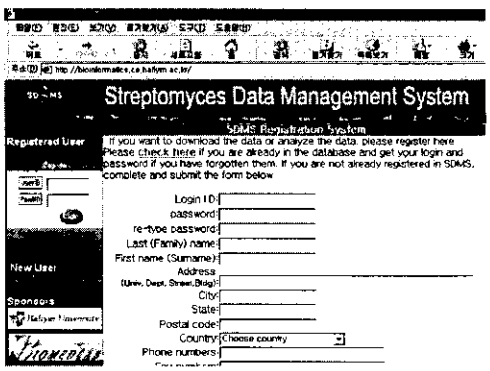

그림 2

3.2.1 News

News는 사용자들에게 새로운 서비스가 추가되었을 때 이와 관련된 정보를 전달하는 기능이다.

3.2.2 Online analysis

Online analysis는 이 시스템의 가장 중요한 기능중의 하나로 사용자들이 비교하기를 원하는 유전자를 입력하여 Streptomyces Database내에 존재하는 유전자들과 비교하여 유사도를 측정할 수 있도록 한다. 이 시스템에서는 유전자 비교분석 도구인 BLAST[2,3]를 사용하였다. 그림 3은 시스템내의 BLAST를 사용할 수 있는 화면이다.

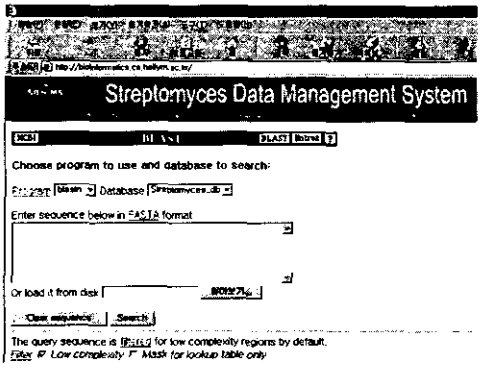

그림 3

유전자 비교분석 도구인 BLAST 프로그램 이외에도 유전자들의 진화관계를 나무의 형태로 나타낼 수 있는 프로그램들을 사용할 수 있도록 하는 기능을 가지고 있다.

3.2.3 Data Download

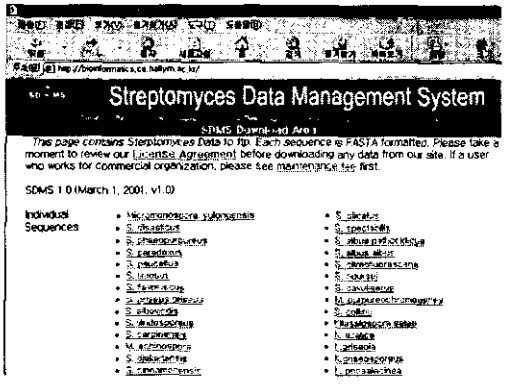

그림 4

Data Download는 이 시스템의 핵심 기능으로 유전자들을 사용자의 시스템으로 다운로드할 수 있는 기능을 가진다. 개 개의 유전자들, 혹은 유전자 집합 전체를 다운로드 할 수도 있다. 또한 모든 유전자들을 정렬(align)하여 손쉽게 유전자들간의 차이를 구별할 수 있도록 한다. 그림 4는 Data Download와 관련된 화면이다.

3.2.4 Contacts

시스템으로 다운로드할 Contacts는 시스템과 관련된 내용을 메일로 문의할 수 있는 기능이다.

3.2.5 Citation

이 시스템에서 제공되는 내용과 관련하여 논문 혹은 기타의 저술에 본 시스템의 내용을 부연할 경우, 사용할 수 있도록 내용을 제공한다.

3.2.6 Link

본 시스템과 관련된 웹 사이트를 링크할 수 있도록 한다. 사용자는 연결된 관련 사이트를 통하여 본 시스템의 내용과 관련된 참고 자료로 이용할 수 있다.

3.2.7 Board

Board는 사용자와 사용자, 사용자와 시스템 운영자간의 의견을 교환할 수 있도록 한다.

3.2.8 Help

Help는 본 시스템의 내용 및 사용법등과 관련된 정보를 제공하기 위한 기능이다.

결론

본 논문에서는 WWW와 데이터베이스를 연동하여 Streptomyces 유전자와 관련된 서비스를 제공하는 유전자 관리 시스템을 설계하여 구현하였다. 이 유전자 관리 시스템의 주요 기능으로는 News, Online analysis, Data Download, Contacts, Citation, Link, Board, Help등이 있다. 사용자는 Streptomyces database내의 유전자들과 비교분석하기를 원하는 유전자를 입력하여 원하는 목적을 달성할 수 있다. 이 관리 시스템의 유전자 데이터를 사용자의 시스템으로 내려 받을 수도 있다.

이 시스템은 국내에서는 매우 드물게 모든 내용이 영어로 작성되어 전세계의 모든 사용자들이 손쉽게 사용할 수 있으며, 국내에서는 유전자들을 효율적으로 관리하는 시스템이 매우 드물어 이 시스템은 국내에서 유사한 시스템을 구축하려 하는 연구자들에게 좋은 선례가 될 것이다.

추후 시스템 서비스의 효율을 높이기 위하여 사용자의 시스템 사용추세와 의견을 참조하여 새로운 기능을 추가할 것이다.

참고 문헌

[1] U.S Congrress. Mapping our genes-the genome projects, how big, how fast? Technical Report OTA-BA-373, Office of Technology Assessment, Washington, D.C., 1998.

[2] Altschul, S. F., Gish, W., Miller, W., Myer, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

[3] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

[4] J. Felsenstein. Numerical methods for inferring evolutionary trees. The Quarterly Reviews of Biology, Vol. 57, No. 4, Dec. 1982.