생명 정보 검색 제공원 비교 및 분석

이승희*, 안후영*, 박영호*
*숙명여자대학교 멀티미디어과학과
e-mail : ariseshine@sookmyung.ac.kr, hyahn85@sookmyung.ac.kr, yhpark@sookmyung.ac.kr

Comparison and Analysis of Bioinformation Resources

Seung Hee Lee*, Hoo Young Ahn*, Young-Ho Park*
*Dept. of Multimedia Science, Sookmyung Women's University

요 약

최근 생명과학의 연구가 복잡하고 다양해짐에 따라 생명과학 기술(BT)과 정보기술(IT)를 결합한 생명정보학의 중요성이 부각되고 있다. 생명정보학을 통해 연구가 진행됨에 따라 생명 정보 데이터의 양이 더욱 방대해지면서 이를 다루는 방법들과 제공원 또한 다양하게 존재하게 되었다. 생명 정보 데이터들은 그 특성과 구성이 복잡하여 그에 맞는 다양한 데이터베이스를 이용하며, 이를 연동하기 위한 통합적인 생명 정보 검색 시스템도 계속해서 연구되고 있다. 본 논문에서는 생명 정보를 다루는 연구 분야에서 연구의 효율성과 확장성을 위한 생명 정보 검색 제공원을 주목하여 그 종류와 시스템 특성에 대해서 설명한다.

1. 서론

최근 생명과학의 연구가 점점 더 다양해지고 복잡해짐에 따라, 생명과학 기술(BT)에 정보기술(IT)을 융합하여 생명 정보를 다루는 생명정보학(Bioinformatics)의 중요성이 높아지고 있다.

생명정보학을 통한 연구가 활발해짐에 따라, 이미 알려진 생물학적 데이터베이스로부터 제공되는 기존 데이터들 이외에도 여러 연구의 실험 데이터나 해석 결과 등으로부터 얻어지는 정보로 인해 데이터의 양이 더욱 방대해지면서 이러한 데이터들을 가공하고 표현 하는 것 역시 중요한 일이 되었다.

생명 정보 데이터들은 각각 그 특성과 구성이 복잡하여 그에 맞는 다양한 데이터베이스들에 저장되어 있다. 현재 국내외 많은 기관들이 연구 목적에 따라 생명 정보 데이터를 저장, 관리하고 검색을 제공하는 데이터베이스를 구축하여 연구에 기여하고 있다. 전 세계적으로 가장 대표적인 생명 정보 데이터베이스 관리 기관으로는 미국의 국립 생물 정보 센터인 NCBI(National Center for Biotechnology)[1], 유럽의 EMBL(European Molecular Biology Laboratory)[2], 일본의 DDBJ(DNA DataBank of Japan)[3] 등을 꼽을 수 있다. 이들은 국제 염기서열 데이터베이스 연합(INSDC)을 통해 거의 동일한 정보를 공유하고 있으며, 데이터의 상호 교환을 하여 데이터의 최신성을 유지하고 있다. 이외에도 단백질 서열이나 3 차 구조 데이터, 신진대사의 경로와 관련한 데이터, 그리고 유전체의 돌연변이나 유전병에 대해서 다루는 등 각각의 연구에 알맞은 데이터를 서비스하는 제공원들이 존재한다.

현재 생명 정보를 분석하는 도구들이 빠르게 발전하고 있으나, 형식이 다른 기존의 데이터들과 신규 데이터들의 통합과 함께 데이터베이스의 표준화가 미비한 실정이다. 본 논문에서는 현존하고 있는 다양한 생명 정보 검색 제공원을 데이터 소스별로 분류하고, 생명 정보 통합 시스템을 살펴본다.

본 논문의 구성은 다음과 같다. 2 장에서는 생명 정보 특성에 따른 생명 정보 검색 제공원을 분류하고, 3 장에서는 생명 정보 검색 제공원에 대한 세부적인 설명을 한다. 마지막으로 4 장에서는 본 논문의 결론을 내린다.

2. 생명 정보 검색 제공원 분류

본 논문에서는 생명 정보를 다루는 다양한 데이터베이스들을 생명 정보 데이터의 특성에 따라 분류하여 비교하였다.

생명 정보 데이터 소스에 따라 분류하면 대표적으로 DNA 서열, 단백질 아미노산 서열, 3 차원 구조, 유전자 발현 정보, 신진대사 경로를 다루는 데이터베이스로 크게 나눌 수 있다. 분류되는 생명 정보 데이터 소스와 그에 따른 데이터베이스는 표 1 과 같다.

<표 1> 생명 정보 데이터 소스 제공원

데이터 소스	데이터베이스
DNA 서열	GenBank (NCBI), EMBL, DDBJ
단백질-아미노산 서열	NBRF, PIR, PRF, Swiss-prot
3 차원 구조	PDB, CS , MMDB
유전자 발현	SMD(Stanford Microarray database)
신진대사 경로	KEGG, WIT, EcoCyc

* 이 논문은 2009 년도 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(No. 2-0905-0012)

이외에도 단백질 구조에 관한 기본적인 데이터베이스를 바탕으로 관련된 motif 를 모아놓은 PROSITE, 잘 보존된 부분에 대한 BLOCKS, PRINTS 와 같은 데이터베이스들도 있다[4]. 또한 OMIM(Online Mendelian Inheritance in Man), HGMD(Human Gene Mutation Database)와 같이 질병, 돌연변이와 관련된 데이터베이스, REBASE 와 같은 제한효소 데이터베이스, 3 차원 폴드 분류, 생체 내의 작용 경로, 생물 종에 따라 연관 유전자를 정리한 데이터베이스들이 존재한다.

3. 본론

본 장에서는 염기 서열 정보와 단백질 구조를 제공하는 대표적인 주요 생명 정보 검색 제공원인 GenBank 와 PDB 에 대해서 설명하고, 이질적인 데이터들을 다루는 데이터베이스들을 연동하고 통합하여 생명 정보 검색을 제공하는 Entrez 와 Ensembl, SRS 을 소개한다.

3.1 GenBank

GenBan[5]는 NCBI 에서 운영하는 유전자 정보 데이터베이스로, Human Genome Project 결과를 포함하여 세계 각지에서 생성된 DNA 염기 서열 정보들을 모은 것이다[6]. mRNA, cDNA, 유전체 DNA, EST 와 같은 서열을 다루며 단백질 정보를 함께 제공하고 있다. Locus, Source, Organism, Features 등의 정보를 담고 있는 GenBank 플랫파일과 여러 생물학 관련 도구에서 보편적으로 쓰이는 기본적 형식인 FASTA 포맷을 다룬다.

3.2 PDB

PDB(Protein Data Bank)[7]는 단백질 3 차원 구조를 저장해 놓은 데이터베이스로, PDB 에 저장되어 있는 데이터들은 X-ray 회절법과 NMR 실험에서부터 나온 실험 데이터들이다. 이 데이터로 인해 단백질이 가지고 있는 정보들과 3 차원 구조를 볼 수 있는 포맷을 제공한다. PDB 플랫파일과 mmCIF 파일, PDBML/XML 파일, 그리고 구조 데이터로부터 유도된 FASTA 형식의 서열 정보를 제공한다.

3.3 Entrez

생명 정보를 검색하는 인터페이스 중 가장 많이 쓰이는 것이 NCBI Entrez[8] 시스템이다. Entrez 는 그 자체가 데이터베이스가 아니라 Entrez 를 통하여 여러 데이터베이스를 검색할 수 있도록 해주는 인터페이스이며, PubMed 논문 정보와 함께 염기 서열과 단백질 서열 아미노산 서열, 단백질 3 차 구조, 염색체 지도 정보 등을 포함하여 한 번의 검색으로 결과를 얻을 수 있다. Entrez 는 하드링크(Hard link)라는 연관관계와 같은 개념으로 통합적으로 정보를 회수한다[9].

3.4 Ensembl

Ensembl[10]은 EMBL-EBI 와 Sanger Institute[11]가 합작하여 만든 시스템이다. 다양한 종(Spieces)의 유전자 정보를 검색할 수 있으며, Ensembl 을 이용하여 서열 데이터 정보와 유전자 예측, 알려진 유전자의 구조 예측이 가능하다. GFF, PSL, BED 와 같은 파일 포맷을 사용한다.

3.5 SRS

SRS(Sequence Retrieval System)는 EBI 에서 개발된 시스템으로, EBI 에서 제공하는 데이터베이스들의 기본 검색 시스템으로 사용되고 있다. 200 여 개의 데이터베이스들을 연결하여 한번에 검색할 수 있게 해주며, 텍스트 기반의 데이터베이스들을 ftp 로 다운로드하여 인덱싱하고 시스템에 통합시킬 수 있다. GenBank 파일 포맷이나, Expasy 에서 제공하는 SwissProt 파일을 SRS 로 포함할 수 있다.

앞에서 소개한 다양한 데이터베이스들은 유전자의 구조를 예측하는 연구에서부터, 서열 비교와 정렬을 통한 유사성 검색, 단백질의 3 차 구조와 기능 예측, 분자간 상호작용, 생체 내의 신진대사 경로, 특정 유전자와 관련된 질병, 유전자의 발현 패턴 등의 수많은 연구들에 사용된다.

4. 결론

본 논문은 생명 정보를 다루는 다양한 데이터들을 저장하고 제공하는 데이터베이스들에 대해서 알아보았다. 생명 정보를 보다 효율적이고 확장된 영역으로 관리하기 위해 위와 같은 제공원들을 효과적으로 이용할 수 있다.

이를 통해 통합 데이터베이스 관리 시스템 개발과 생명 정보 데이터마이닝 기법, 염기 서열로부터 단백질의 구조와 기능을 예측하는 방법 등의 연구가 활발이 진행되고 있으며, 이질적인 파일 형식을 다루는 데이터베이스 상호간의 연동성과 복잡한 분석이나 검색의 정확성을 높이는 등의 연구가 중요시될 것으로 예상된다.

참고문헌

[1] NCBI, http://www.ncbi.nlm.nih.gov/

[2] EMBL, http://www.ebi.ac.uk/embl/

[3] DDBJ, http://www.ddbj.nig.ac.jp/index-e.html

[4] 최현화, 채미옥, 이미영, “생물학 데이터베이스 및 웹 기반 통합검색 시스템의 현황 및 전망,”. *한국전자통신연구원 주간기술동향*, Vol. 1074, 2002

[5] GenBank, http://www.ncbi.nlm.nih.gov/Genbank/

[6] DA Benson, MS Boguski, DJ Lipman, J Ostell, BF Ouellette, BA Rapp and DL Wheeler, “GenBank,”. Nucleic Acids Research, Vol.27, No.1, pp.12~17, 1999

[7] PDB, http://www.rcsb.org/pdb/home/home.do

[8] Entrez, http://www.ncbi.nlm.nih.gov/Entrez/

[9] 안드레아스 박스바니스, 프란시스 올레뜨, “생물정보학, 유전자와 단백질 분석에 관한 실용 지침서,”. 제2판, 월드사이언스, 2003

[10] Ensembl, http://www.ensembl.org/index.html

[11] Sanger Institute, http://www.sanger.ac.uk/

[12] 이성훈, 강성후, 김현철, “SRS를 이용한 바이오데이터베이스의 통합 검색법,”. *분자세포생물학뉴스*, Vol.16, No.1, pp.14~19, 2004