제9절 유전자 데이터베이스

1. 유전자 데이터베이스

유전자는 단백질로 번역되어 세포 내에서 특정한 생물학적 기능을 수행한다. 한 유전자와 관련이 있는 정보에는 유전자의 생물학적 출처, 유전자의 발현여부를 결정하는 조절단백질의 특성,

유전자의 산물, 그 산물에 관여하는 대사회로의 종류나 작용하는 기질까지 포함된다. 유전자와 연관 정보를 담은 데이터들은 인간 유전체 프로젝트(human genome project)가 진행됨에 따라 엄청난 수로 급증해왔다. 2002년 8월 현재 세계적으로 635개의 유전체 프로젝트가 진행 중이고, 이 중 이미 완료가 된 것이 100개이고, 나머지는 현재 진행 중이다 (http://igweb.integratedgenomics.com/GOLD/). 이러한 연구로부터 얻을 수 있는 1차 적인 데이터는 서열 데이터이며 불과 몇 안되는 생물체에서 이루어지고 있다고는 하지만 그 양은 매우 크기 때문에 컴퓨터와 같은 도구를 이용해 데이터베이스를 구축하여 다음 단계의 연구를 가능하게 한다. 2001년까지 공개된 세계의 생물정보 관련 데이터베이스는 335개(Nucleic Acids Research 2002 30:1-12)이며, 이 데이터베이스들은 개별적 특정 형태인 DNA, 단백질 서열 정보뿐만이 아니라, 3차원 구조데이터, 유전체 데이터, 생화학적 경로 데이터, 유전자 발현 데이터 등을 포함하고 있다. 서열 데이터를 한데 모아 특정한 데이터베이스 시스템을 구축하면 여러 가지 전산학적 기법을 응용하여 유전체들의 상관관계나 연관성을 파악해 유전자들의 진화과정 및 연관 정보를 유추할 수 있게 된다. 1990년대 이후 인터넷의 보급과 컴퓨터 하드웨어 발전으로 사용자는 동일한 인터페이스 내에서 공개된 데이터베이스에 쉽게 접근하여 문자로 표현된 서열들로부터 잠재적으로 의미 있는 생물정보를 유추할 수 있다.

유전자 데이터베이스 중에서 단백질에 대한 3차원 구조 데이터를 다룬 PDB(Protein Data Bank)는 구조생물정보학 협력기구(RCSB, Research Collaboratory for Structural Bioinformatics)에 의해 운영되고 있다. 이용자는 PDB identifer, Searchlite, SearchFields 인터페이스를 사용하여 검색할 수 있고, 단백질과 탄수화물, 분자합성물의 구조를 RasMol이나 Chime와 같은 플러그인 브라우저로 볼 수 있다.

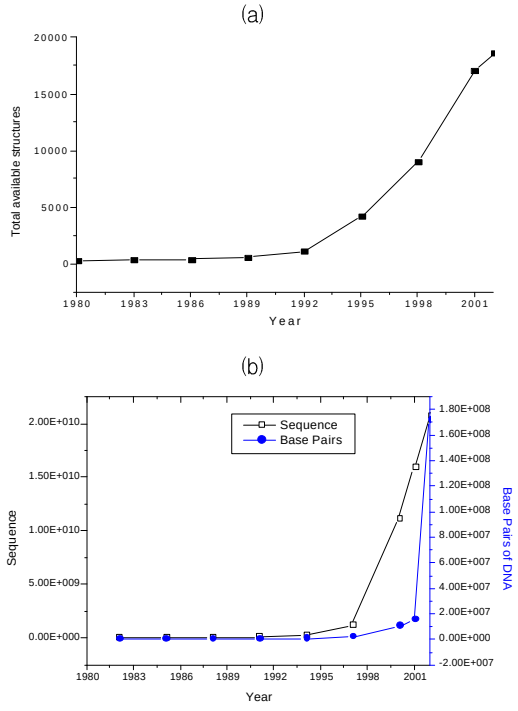
[표 4-10-19] 주요한 유전자 데이터베이스 주소

데이터베이스	기 관	주 소
MEDLINE	국립 의학도서관	www.nlm.nih.gov
GenBank	국립 생명공학 정보센터	www.ncbi.nlm.nih.gov
EMBL	유럽 생물정보학 연구소	www.ebi.ac.uk
DDBJ	일본 국립 유전학 연구소	www.ddbj.nig.ac.jp
SWISS-PROT	스위스 생물정보학 연구소	www.expasy.ch
PIR	국립 생명의학 연구재단	www.nbrf.georgetown.edu
PRF	일본 단백질 연구재단	www.prf.or.jp
PDB	구조생명정보학 연구공동체	www.rcsb.org/pdb/
CSD	캠브리지 결정학 데이터 센터	www.ccdc.cam.ac.uk

자료 : Post-genome informatics, Minoru Kanehisa, 2000.

NCBI GenBank는 전 세계에서 산출되는 거의 모든 종류의 서열 데이터가 저장되어 있는 세계 최대의 생물정보 데이터베이스이다. GenBank는 cDNA, EST(Expressed Sequence Tag), SNP(Single Nucleotide Polymorphism), STS(Sequence Tagged Sites) 등에 이르는 광범위한 유전체 정보를 포함한다. 또한 GenBank는 데이터베이스의 상호 운용성을 증진하기 위해 표준 관계형 데이터베이스 양식 ASN.1 포맷을 활용하고 있고, 서식 기반 인터페이스를 통하여 사용자가 입력한 데이터를 FASTA 양식으로 저장한다. 2002년 6월 Genbank에서 배포한 17,471,000 서열 내에 염기쌍이 20,649,000,000이 존재한다(GenBank. Nuclic Acids Reserach 2002 Jan 1;30(1):17-20).

이 외에도 NCBI의 Entrez (http://ncbi.nlm.nih.gov:80/entrez/)는 식물 동물 모델 시스템의 유전체들을 부분적이거나 완전하게 자료를 제공하는 유전체 데이터베이스와 WIT(http://wit.mcs.gov/WIT2/)와 KEGG(http://genome.ad.jp/kegg/) 등과 같은 분자들의 네트워크를 통한 생물학적 활동을 다룬 경로 데이터베이스 시스



[그림 4-10-43] 생물학적 데이터베이스 증가 추세

자료 : (a) PDB의 분자구조 증가 추세 배포문서, (b) Genbank의 염기서열의 양 증가 추세 배포문서

템이 있다.

2. 유전자 데이터베이스를 이용한 유용한 정보추출

가. 데이터베이스 검색 프로그램을 이용한 유전자 기능예측

"서로 다른 분자의 서열간에 상동성(homology) 혹은 유사성(similarity)이 존재한다면, 유사한 3차원적 구조나 비슷한 기능을 지닌다"라는 가설은 실험적으로 증명되어야 인정받을 수 있다. 그러나 유전자의 기능은 진화적인 관계나 물리 화학적 제약이 있으므로 새로운 서열을 확보하면 다른 서열데이터정보를 이용하여 상동성 검색과 단백질 코딩영역, 엑손(exon)과 인트론(intron)의 경계 등을 예측할 수도 있다. 분자의 서열로부터 유전자 기능을 예측하기 위한 가장 간단한 접근은 두 서열을 임의적으로 놓고 생물학적으로 의미가 있는 점수 함수(Score Function)를 최적화하여 최적의 정렬을 하는 것이다.

[표 4-10-20] 서열비교 하는 웹서버

LFASTA at PBIL	http://pbil.univ-lyon1.fr/lfasta.html
BLAST twosequences at NCBI	http://www.ncbi.nlm.nih.gov/grof/bl2.html
LALIGN at CRBM	http://www2.igh.cnrs.fr/bin/lalign-guess.cgi
SIM, GAP, NAP, LAP	http://genome.cs.mtu.edu/align/align.html
PSI-BLAST	http://www.ncbi.nlm.nih.gov/cgi-bin/blast/psiblast.cgi
MSA at IBC	http://www.ibc.wustl.edu/ibc/msa.html
ClustalW at EBI	http://www2.ebi.ac.uk/clustalw/

자료 : Bioinformatics: Sequence, Structure, and databanks, D.Higgins, W.Taylor, 2000

BLAST는 우선 질의서열(Query Sequence)과 정렬을 이루었을 때 임계치(Threshold)이상의 점수를 기록하는 문자 블록의 목록을 만든다. 다음 단계로 미리 계산이 된 표를 이용하여 질의 서열과 서열 데이터베이스간의 서열 유사성이 없어질 때까지 유사영역을 횡으로 늘려 가는 방법으로 국부정렬(Local Alignment)을 수행한다. BLAST는 국부정렬이 HSP (High-scoring Segment Pairs)순서로 정렬하기 때문에, 유전자의 선두에서 말미까지의 위치순서로 정렬하는 것은 아니다.

FASTA는 해쉬 테이블(Hash Table)로 불리는 검색 테이블을 사용하여 질의 서열과 서열 데이터베이스의 양쪽에 들어있는 짧은 서열의 일치되는 위치를 찾는다. FASTA는 치환이나 삽입에 기인하는 갭(Gap)을 허용하여 조합해 가면서 최적의 국부정렬을 수행한다. FASTA는 BLAST와 마찬가지로 두 개의 비교적 긴 서열에서 상동성이 존재하는 짧은 부분의 탐색이다. 문자 블록 단위 내에서 서열의 상동성은 적지만, 생물학적 기능이 유사한 단백질 서열일 수 있기 때문에, 두 서열만을 비교하는 것보다는 여러 서열을 비교하는 것이 서열로부터 유전자 기능 예측하는 데에 더 일반적인 접근이다.

ClustalW에서는 계통발생학적 분석에 기반을

두었으며, 모든 서열에 대한 거리함수를 계산한 후, 두 서열 사이의 유사성을 반영하는 유도 트리(Guide Tree)를 만든다. ClustalW는 서열 가운데 가장 관계가 가까운 짝부터 트리의 가장 바깥쪽 가지까지 동적 프로그래밍(Dynamic Programing)으로 정렬한 후에 각각의 새로운 정렬을 분석하여 두 서열 사이, 서열과 그룹들 사이, 두 그룹들 사이의 정렬을 전진적 다중 정렬을 수행해서 서열 전체를 정렬한다.

다중 정렬을 하면 서열 간 잘 보존된 영역과 분산되어 있는 영역이 있는데, 서열에서 보존된 영역은 진화의 과정에서 변화하기 어려워서 근원이 된 서열군이 갖는 기능을 지닌다. 단백질에서 구조적 기능적 특징에 해당하는 서열의 부분적인 보존영역이나 서열이 공유하는 짧은 서열의 패턴을 모티프(Motif)라 한다. 짝 정렬에 모티프정보를 활용하여 최적화 할 수 있는 다중서열정렬방법은 다수의 연관된 서열의 짝 정렬 결과와 단백질 데이터베이스내의 유전자 패밀리의 모든 구성원에서 서열 상동성이 있는 결과들을 결합시켜 단백질의 유사성, 모티프의 발굴, 유전적 상관관계를 알아낸다.

짝 정렬(Pairwise Sequence Alignment)에

[표 4-10-21] 서열패턴을 이용한 다중정렬 하는 웹서버

Block-based global multiple alignment	
DCA at BiBiServ	http://bibiserv.techfak.uni-bielefeld.de/dca
DIALIGN2 at BiBiServ	http://bibiserv.TechFak.uni-biblefeld.de/dialign
ITERALIGN at Stanford	http://giotto.stanford.edu/luciano/iteralign.html
Motif-based local multiple alignment	
MEME at SDSC	http://www.sdsc.edu/MEME/
MEME at Pasteur	http://bioweb.pasterur.fr/sequanal/motif/meme/
BLOCK Marker at FHCRC	http://blocks.fhcrc.org/blockmkr/make_blocks.html
PIMA II at BMERC	http://bmerc-www.bu.edu/protein-seq/pimall-new.html

자료 : Bioinformatics: Sequence, Structure, and databanks, D.Higgins, W.Taylor, 2000

특정 도메인(Domain)이나 모티프(Motif)를 활용하는 PSI-BLAST(Position-Specific Iterated BLAST)는 단일 서열로 시작하여 갭을 허용하는 국부적인 다중 정렬을 사용한다. PHI-BLAST (Pattern-Hit Iterated BLAST)는 모티프 대신 입력한 질의 서열과 단백질 서열 데이터베이스 내에서 선택한 서열간의 패턴을 이용하여 다중 정렬한다.

Fred Hutchinson Cancer Research Center의 서비스인 Blocks는 근원이 되는 정렬 부위(Seed Alignment)를 찾기 위하여 서열에 있는 3개씩(Triplet) 아미노산 공간을 검색하면서 최대정렬부위를 찾아 확장하는 모티프 검출방법의 조합으로 만들어진 국부 서열정렬 시스템 이다. 이 데이터베이스는 상동성을 가지는 단백질 서열로부터 새로운 모티프를 찾는데 유용하다.

스위스 생명정보공학 연구소(SIB, Swiss Institute of Bioinfomatics)에서 운영하는 PROSITE는 다중정렬의 보존부위의 아미노산 서열을 대표하는 패턴으로 만들어서 그 패턴이 새로운 서열에 존재하는지 탐색하는 단백질 패턴 데이터베이스이고, PRINTS는 PROSITE와 유사한 단백질 모티프 데이터베이스이지만, 패턴 대신에 지문(Fingerprints)을 사용한다.

Pfam은 서열정렬 결과로부터 아미노산이 얼마나 출현하는지의 빈도를 나타내는 행렬을 만들고, HMM(Hidden Markov Model)를 이용하여 그것과 미지의 서열이 적합한지 아닌지를 맞추어서 데이터베이스로부터 한 종류의 전체정렬을 만든다.

나. 단백질 입체구조에 근거한 유전자 기능 예측

단백질은 아미노산 배열에 의해 1차 적인 구조가 이미 결정되고, 단백질 접힘(Folding)을 통해 고유의 기능을 수행 할 수 있는 구조를 이룬다.

이미 알려진 단백질과 서열을 비교함으로써 생물학적 기능이나 계통발생학적 진화에 대한 정보를 얻을 수 있지만, 단백질의 기능을 이해하기 위해서는 서열분석 보다는 3차원 구조를 분석하는 것이 더 효과적이다. 단백질 기능에 대한

구조의 중요성을 인지하여 세계적으로 모든 접힘 구조를 NMR 또는 X선 결정해석으로 밝히는 프로젝트를 진행하고 있다.

단백질 2차 구조의 패턴은 단백질 접힘 구조의 분류를 예측할 수 있게 하기 때문에 단백질 3차원 구조 예측의 첫 단계로서 가치가 있다. 단백질 2차 구조 예측은 그 패턴을 이용하여 단백질 아미노산 서열의 국부적인 구조인 알파 나선 구조, 베타 구조, 코일구조로 단백질을 분류한다.

단백질 2차 구조 예측을 할 수 있는 방법에는 이미 알려진 단백질의 아미노산 서열에서 미지의 서열과 유사성을 검색하는 서열 정렬을 기반으로 하는 방법과 미지의 서열만을 가지고 구조를 예측하는 단일 서열기반 방법이 있다.

[표 4-10-22] 단백질 2차 구조 예측 프로그램의 웹 서버

GOR4	http://absalpha.drt.nih.gov:8008/gor.html
PHD	http://dodo.cpmc.columbia.edu/predictprotein/
Pred2ary	http://yuri.harvard.edu/~jmc/2ary.html
PREDATOR	http://embl-heidelberg.de/cgi/predator_serv.pl
DSC	http://bonsai.lif.icnet.uk/bmm/dsc/dsc_read_align.html
Zpred	http://kestrel.ludwig.uk.ac.uk/zpred.html
Jpred	http://barton.ebi.ac.uk/servers/jpred.html
COILS2	http://www.isrec.isb-sib.ch/coils/COIL_doc.html

자료 : Bioinformatics: Sequence, Structure, and databanks, D.Higgins, W.Taylor, 2000

[표 4-10-22]에서 보여주는 단백질 구조 분석 시스템에서 PHD 프로그램은 단백질의 서열 길이, 아미노산의 출현 빈도수 등과 단백질 서열의 특성을 기반으로 신경망 예측방법의 결과를 혼합해 사용하고, Jpred는 다중 서열정보 이용한다.

PSA는 2차 구조를 예측하기 위해 마르코프 모델(Markov model)을 이용하고, PREDATOR는 아미노산의 수소결합특성에 대한 다중서열 정보를 함께 처리한다.

광합성, 뉴런의 흥분작용, 면역반응, 한 세포에서 다른 세포로의 신호 전달 등을 하는 막횡단 나선(Transmembrane Helices)구조는 결정화가 되기 어렵고 수용액 상태에서 안정하지 못해서 X선 결정해석 또는 NMR등의 실험적인 방법으로 구조 예측이 어렵다. 이러한 구조를 예측하기 위해 사용하는 방법은 단백질의 아미노산 서열에서 세포막을 관통할 수 있는 아미노산 17-25개정도의 단위로 나선 구조 접힘 구조를 만들어서 소수성을 계산하여 세포막에서 존재할 수 있는 가능성을 확인하는 것이다. 이 방법의 일련의 과정은 단백질의 2차 구조를 예측하는 과정과 관련이 있다. 막횡단 나선구조를 예측할 수 있는 웹서버는 TopPred(http://www.sbc.su.se/~erikw/toppred2), TMHMM(http://www.cbs.dtu.dk/services/TMHM/) 등이 있다.

단백질의 서열만을 가지고 직접적으로 3차원 구조를 정확하게 예측하는 것은 매우 어려운 일이다. 기존의 구조 데이터베이스에서 물리 화학적 환경 변수를 추출하여 구조가 아직 밝혀지지 않은 단백질의 아미노산 서열에서 최적화된 구조를 예측하는 지식기반 방법으로 접근하고 있다. 기능을 알고 있는 단백질들의 접힘 형태 등을 따로 모아 기능이 알려지지 않은 단백질의 아미노산의 접힘을 입력하여 어떤 접힘 형태와 유사한지를 알 수 있다면, 단백질의 기능을 예측할 수 있다.

단백질의 3차원 구조 예측 방법에는 단백질의 문자서열 위에 있는 아미노산의 화학적인 특성을 정량적인 분석하여 단백질 구조에 대한 3차원 정보를 3차원 대 1차원으로 일일이 대응관계로 나타내는 방법, 그리고 구조를 이미 알고 있는 단백질에서 모든 아미노산 잔기의 조합에서 출현빈도를 헤아려서 에너지 계수를 결정하여 어느 아미노산 배열이 어느 자리에 접하는지 정량적인 에너지로 측정하는 방법들이 있다.

접힘 유사성을 이용하여 단백질 분류하는 SCOP(Structural Classification of Proteins, http://scop.mrc-lmb.cama.ac.uk/scop)데이터베이스는 단백질의 2차 구조 특성에 따라 알파 나선구조, 베타구조, 알파나선과 베타구조의 복합된 형태, 금속이온을 지닌 작은 단백질 그룹으로 나누고 있고, 또한 접힘 단계(Fold Level)로 단백질의 위상기하학(Topology)적인 방법에 따른 분류와 알려진 기능과 연관된 도메인에 따른

분류도 포함하고 있다. SCOP에서 단백질의 3차원 구조는 PDB처럼 RasMol이나 Chime 플러그인 브라우저를 이용하여 볼 수 있다. 단백질의 접힘 구조를 이용하여 단백질을 분류하는 방법에는 CATH(http://www.biochem.ac.uk/bsm/cath_new)데이터베이스와 MMDB(http://www.ncbi.nlm.nih.gov/) 등이 있다.

3차원 구조가 밝혀진 단백질과 어느 정도 아미노산 배열의 일치도가 있는 경우에는 구조를 알고 있는 단백질을 주형으로 삼아 새로운 단백질의 구조를 예측하는 상동성 모델링을 수행한다. 상동성 모델링은 주형 단백질과 유사한 기질이 결합하는 것인지 다른 기질이 결합하는 것인지에 대한기질 특이성도 예측가능 하지만, 아미노산의 배열간의 일치도가 낮으면 모델의 정확도도 또한 낮아진다.

SWISS-Model(www.expasy.ch/swissmod/SWISS-MODEL.html)은 웹 서버를 이용하여 상동성 모델링을 할 수 있다.

상동성 모델링처럼 구조를 알고 있는 단백질의 주형을 사용하는 것이 아니라, 아미노산 배열 정보만으로 단백질의 입체구조를 예측하는 순이론적(ab initio)법은 이론과 컴퓨터의 방대한 계산력을 요구하기 때문에 작은 단백질에서만 적용할 수 있다.

단백질의 기능적인 부분은 국부적인 부분에서 결정되어지기 때문에 그 부분에 유사성을 지니는 단백질의 기능에 관여하는 모티프 부분이 유사한 구조를 지닌다면, 단백질의 기능도 유사할 것이다. 모티프를 이용하여 단백질의 전체 접힘이 아니라 부분 입체구조를 조사하여 기능을 예측할 수 있다.

3. 유전자 데이터베이스 전망

[그림 4-10-43]에서 이미 보여준 것과 같이 유전자를 포함한 생물학적 데이터베이스는 급격하게 증가하고 있으며, 이러한 경향은 앞으로도 계속될 것으로 보인다. 이와 같이 급속한 데이터의 증가 경향은 생물 정보 데이터베이스 이외의 시스템에서는 찾아 볼 수 없는 현상이다. 생물정보 데이터베이스를 합리적으로 저장하고 관리하는 것은 다른 데이터베이스에서 보다 더 매우 중요하다 할 수 있다. 또한 이러한 일차 데이터베이스를 목적에 따라 이차적으로 가공하여 새로운 정보를 만들어 내는 것은 새로운 과학을 탄생시킬 정도로 중요하다. 따라서 우리나라의 경우도 위에서 언급되어진 1차원적 데이터베이스를 우선적으로 구축하여 우리나라 과학자들이 빠르고 편리하게 이용할 수 있도록 하여야 할 것이며, 이를 가공하여 저장하는 2차 데이터베이스도 목적에 따라 구축하는 것이 매우 필요하다 할 것이다. 이러한 작업은 단순히 한 개의 단체가 수행할 수 있는 일이 아니므로, 많은 연구단체들이 유기적으로 정보를 교환하여 필요한 데이터베이스를 구축하고 구축된 데이터베이스를 공용해야한다.

4. 새로운 데이터베이스 구축 기술 개발의 필요성

GenBank를 비롯한 각종 생물정보 데이터베이스의 양은 지난 수년간에 비해 매우 증가하여 왔으며, 앞으로는 어떠한 속도로 더 빠르게 증가할지조차 예측하기 힘들다. 현재까지 대부분의 유전체/단백체 데이터베이스들은 데이터의 입출력과 검색을 위해 Oracle이나 MYSQL 등의 DBMS 시스템을 이용하는 것이 일반적이다. 물론, Arabidopsis 데이터베이스처럼 자체적으로 개발(ACEDB)하여 구축하는 경우도 존재하지만 상용 데이터베이스 관리 시스템을 이용하는 것이 보편화되어 있다. 그러나 데이터의 양뿐만 아니라 증가 속도가 매우 빠른 속도로 늘어나고 있음을 고려할 때 새로운 형태의 데이터 관리 및 검색 기능을 수행할 수 있는 시스템 개발이 이루어져야 한다. 한 예로, GenBank를 재구축하기 위해 Oracle과 같은 기존의 DBMS를 이용한 시도는 국내에서 몇 차례 있었으나 그 효율성이 저하되어 크게 이용하지 못하였다. 그러나 최근 한국과학기술정보연구원(KISTI)은 자

체적으로 개발한 검색 엔진인 KRISTAL을 응용한 BIO-KRISTAL을 통해 현재 약 1천 8백만 건에 달하는 GenBank 검색 시스템을 재구축하여 빠른 속도로 데이터 검색을 수행할 수 있도록 하는데 성공하였다. 이러한 기술을 토대로 유사하거나 서로 관련성이 높은 수많은 데이터베이스들을 통합한 "생물정보 통합 데이터베이스"의 구축이 사실상 이루어질 수 있게 되었다. 검색엔진을 이용한 유전체/단백체 데이터베이스 구축은 정형화되어 있는 틀을 갖는 기존의 DBMS 시스템에 비해 조건에 따른 응용성이 높게 평가되고 있다.