약제학회지(2004), 제 34 권 제 1 호
J. Kor. Pharm. Sci., Vol. 34, No. 1, 73-79 (2004)

바이오데이터베이스와 도구를 활용한 바이오인포매틱스의 동향

임달혁*,** · 전수경** · 박완규* · 이영주*,†
*세종대학교 생명공학과, **(주) 비츠 코리아
(2004년 1월 6일 접수 · 2004년 2월 6일 승인)

Current Status of Bioinformatics on Bio-databases and its Tools

DalHyuk Im*,, SueKyoung Jeon**, WanKyu Park* and YoungJoo Lee*,†**

**College of Engineering, Institute of Biotechnology, Department of Bioscience and Biotechnology, Sejong University, Seoul 143-747, Korea*
***BitzKorea Inc., Gangnam-Gu, Seoul 891-39, Korea*
(Received January 6, 2004 · Accepted February 6, 2004)

ABSTRACT–The union of information-technology and biology presents great possibilities to both applications of bio-information and development of science and technology. Also, meaningful analysis of bio-information brings about a new innovation in the field of bio-market with the advent and growth of bioinformatics. Hence, bioinformatics is the most important aspect for establishing a science-technology-oriented society in the 21st century. This article provides trends in current state of bioinformatics. Technological development of bioinformatics for the rapid growth of bio-industry means that using bioinformatics, a biologist can process and store enormous amount of data such as current Human Genome Project and future data in the field of biology. We have mainly looked at the trends of bio-information, databases and mining tools that are generally used, and strategies and directions for the future.

Key words–Bioinformatics, Biodatabase, Mining tool

생명 현상의 유지와 변화를 이해하기 위한 인간의 노력은 생명과학의 발전과 더불어 정보기술 분석의 발전을 가져왔다. 이로 인해 31억 쌍에 달하는 DNA 염기정보와 기능성 유전자의 정보 분석은 바이오인포매틱스의 기술과 자원을 바탕으로 하여 점진적인 진행이 이루어지고 있다.[1] 지난 20세기 생명과학 분야의 최대 발견이라 해도 과언이 아닌 제임스 왓슨과 프랜시스 크릭의 DNA 이중 나선 구조의 규명은 수학자인 존 그리피스와 어윈 샤르가프의 계산적인 접근 방식이 없었다면 완전할 수 없었던 것과 같이 21세기에 인류가 이루어야 할 가장 큰 업적의 밑거름이 될 수 있는 것이 바이오와 정보기술의 만남인 것이다.[2] 이와 같이 생명과학 분야에서 바이오 산업과 정보기술분야의 분석 산업을 분리하여 생각할 수 없으며, 각 처에서 연구되고 있는 다양하고 연관된 실험결과 데이터를 모두 한 곳으로 집결시켜 관리하는 데이터베이스 단위의 정보 공유화에 노력하고 있다.[3] 올해는 DNA 이중나선 구조가 규명된 지 50주년이 되는 해이자 분자 생물학의 획기적인 혁신을 가져온 유전자 재조합 기술, PCR(polymerase chain reaction) 기술 또한 각각 30주년, 20주년이 되는 해로, 이는 인간 유전체 프로젝트를 지난 4월에 완성시킨 성과와 더불어 인류과학에 있어 매우 의미있는 일이라 할 수 있다.[1,4] 이러한 생명과학 역사의 흐름 속에 바이오인포매틱스의 시장 역시 자연스러운 성장 발전을 거듭하고 있다. 이에 본 리뷰에서는 바이오인포매틱스의 정의와 응용되고 있는 분야에 대해 알아보고, 현재 가장 많이 사용되고 있는 데이터베이스와 이를 활용하기 위한 도구들을 소개함으로써 발전하는 바이오인포매틱스의 동향을 말하고자 한다.

†본 논문에 관한 문의는 이 저자에게로
Tel : 02)3408-3766, E-mail : yjlee@sejong.ac.kr

바이오인포매틱스의 동향

바이오인포매틱스의 정의 및 응용분야

1965년 생물학적 데이터를 바탕으로 컴퓨터를 활용한 연구결과들을 'computational biology'라는 용어로 불리기 시작하여 1968년 Rybak에 의해 'bio-informatics'라는 말이 만들어지게 되고, 그 후로 10년이 지난 1978년에 '생물학적 문제를 다루는 학문'으로 규정 되었다.[5] 그럼 '바이오인포매틱스는 대체 무엇인가. 컴퓨터를 사용하는 모든 생물학 작업이 바이오인포매틱스인가'라는 질문에 바이오인포매틱스는 학문의 개념보다 'biology'를 위해 사용되는 수많은 도구에 더

가깝고, 'biology'는 학문이되 바이오인포매틱스는 그에 관한 연구라는 의견도 있다.[6,7] 이는 다시 말해 바이오인포매틱스는 역시 'biology'이며, high-throughput biology, integrative biology라는 것으로 생물학의 진정한 결실을 위해 필요한 도구적인 개념의 학문이라는 것이다.

바이오인포매틱스는 왓슨과 크릭에 의해 DNA의 이중나선 구조가 밝혀지고,[1] 단백질의 아미노산 서열이 결정되면서 시작되었다고 할 수 있다. 이들 단백질에 대한 정보의 데이터베이스가 구축되며, 미국 국립 생명의학연구재단(national biomedical research foundation, NBRF)이 단백질 서열 데이터베이스를 만들게 됨으로써 미 국립보건원(national institutes of health, NIH)의 지원하에 단백질정보센터가 설립되었다. 또한 DNA 염기서열 데이터를 분석할 수 있는 기술이 개발되고 생명과학에 있어서 정보기술의 도입이 본격화 되기 시작하였다. 미 국립 로스아라모스 연구소의 GenBank와 유럽 분자생물연구소인 EMBL데이터베이스가 1982년 공식적으로 출범하였고, 2년 뒤 일본의 DNA 데이터 은행인 DDBJ가 결성 되었다. 이들 세 기관은 정보 네트워크의 구축을 위해 국제염기서열 데이터베이스 협력기구를 결성한 상태이며, GenBank는 1992년을 기점으로 국립 생명공학 정보센터(national center for biotechnology information, NCBI) 산하로 이관되어 현재 GenBank, RefSeq, PDB를 포함한 2003년 12월 Entrez 염기서열 데이터는 2.0×10^{10} (20,197,497,568)을 넘는 상태이다.[8,9] 이러한 바이오인포매틱스 데이터를 이용한 분석과 가공의 과정은 4가지 분야로 나뉠 수 있다. 서열 자동 분석기를 이용한 염기서열 분석, 단백질의 아미노산 서열 발현부분을 구분 짓고 3차원 구조를 파악하는 구조분석, 유전자와 단백질의 기능을 분석하는 기능해석 그리고 유전자의 상호작용과 대사경로를 알아 내는 대사경로 분석이 그것이다.[8,10] 또한 유전체학에서 구분되는 분야의 바이오인포매틱스의 방법은 개별 유전체학, 기능 유전체학, 비교 유전체학, 구조 유전체학으로 나뉘어 구분되어지며,[11] 데이터베이스를 기반으로 하여 유전체 데이터들의 네트워크를 만들어 상호적으로 연구하는데 이용된다.

생물학의 모든 분야에 응용되는 'biology 도구'인 바이오인포매틱스의 분야는 전산학, 수학, 통계학 등의 접목으로 sequence comparison의 유사성 검색 및 정열, 데이터의 의미 해석표시(annotation), 유전자의 조절, 발현부위 등을 찾는 유전자 모델링, 분자구조연구, 유전자 및 단백질 상호작용 연구 등의 분야가 있고, 이를 응용하여 발전시킬 수 있는 응용분야가 있다.[12,13,14] 바이오 데이터를 기반으로 정보기술을 활용하는 바이오인포매틱스의 응용분야는 생명체의 막대한 데이터를 수집, 저장, 분석 및 통합하는 것으로 생명정보의 가공된 기술들을 생명과학으로 포함되는 모든 분야에 적용하게 된다. 이는 데이터베이스로 구축되어 의학, 제약, 농업, 식품, 화학, 환경, 진단 및 생물학 등의 분야에 활용되고, 이로 인해 만들어 질 부가가치는 막대할 것으로 예측된다. 2000년 Nature Biotechnology의 자료에 따르면 바이오인포매틱스의 활용분야는 제약 21%, 지노믹스 12%, 농업 11% 그 외 생물학 7%, 진단 6%, 화학 5% 환경 2%, 기타 36% 등의 순으로 나타났다.[12,15] 제약분야에서 바이오인포매틱스에 대한 활용은 신약개발을 목표로 하여 생명기술인 BT(Bio-Technology) 부분과 정보기술의 IT(Information-Technology) 부분, 분자 설계에 기반한 NT(Nano-Technology)를 바탕으로 진행된다. 이러한 기술들을 통하여 신약 후보 물질을 개발하기 위한 target 유전자 예측을 실시하고, 유전자에서 생성되는 단백질 분석의 3차원적 구조와 기능을 연구한다. 또한 고감도로 분리 정제된 후보 약물에 대한 적합성 검토와 약동역학에 대한 성질 평가 및 다양한 조건 분석에 있어서 바이오인포매틱스 도구(S/W)를 활용한 화합물 라이브러리를 사용하게 된다. 또한 Vitual-Cell 또는 Electronic-Cell을 통한 계산적인 소프트웨어를 통해 가상 세포를 구현하는 방식으로 임상세포 반응 수집에 관한 수많은 임상 실험을 최소화 함으로써, 신약 개발에 활용될 수 있도록 바이오인포매틱스에 관한 역할이 확대되고 있다.

데이터베이스 및 마이닝 툴

바이오인포매틱스의 데이터를 얻는 방식은 대량화된 분석 기술로 인해 빠르고 다양하게 발전하고 있다. 대량의 고속분석 기기들로 인해 데이터의 대량 생산 및 비용절감이 가능해 졌으며, 자동 서열 분석기기를 비롯하여 대량 유전자의 발현패턴을 동시에 볼 수 있는 microarray, 세포생리학적 분석을 시각화하여 할 수 있는 image analyzer, mass spectroscope 등이 이에 속한다.[7,16,17,18] 이러한 방식으로 얻어진 데이터는 무엇보다 분석처리가 중요하며, 사람의 혈액이나 조직에서 분리한 표적 세포의 DNA를 비교하는 염기서열 분석의 경우 목적에 맞는 적절한 소프트웨어를 이용한 통합화된 분석 방법은 연구자가 원하는 결과를 얻을 수 있는 빠른 처리 방법의 예로 볼 수 있다. 뿐만 아니라 데이터를 정리 및 가공하여 상호 연결하는데 있어서는 처리 도구와 더불어 데이터베이스가 중요한 구성요소이다. 바이오인포매틱스에 사용되는 데이터베이스의 특징은 이를 잘 활용할 수 있도록 하는 검색시스템이 데이터베이스의 지원 도구로 사용되는 것이다(Table I). 따라서 데이터베이스의 운영기관에 따라 다른 시스템 체제로 미국(NCBI), 유럽(EBI), 일본(DDBJ) 정부가 그들 각 센터의 활발한 운영을 지원하고 있고, 이곳의 데이

Table I–*Internet Resources for Biodatabases (Ref. 3)*

Database	URL
Protein sequence (primary)	
SWISS-PROT	www.expasy.ch/sprot/sprot-top.html
PIR-International	www.mips.biochem.mpg.de/proj/protseqdb
Protein sequence (composite)	
OWL	www.bioinf.man.ac.uk/dbbrowser/OWL
NRDB	www.ncbi.nlm.nih.gov/enterz/query.fcgi?db=Protein
Protein sequence (secondary)	
PROSITE	www.expasy.ch/prosite
PRINTS	www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html
Pfam	www.sanger.ac.uk/Pfam
Macromolecular structures	
Protein Data Bank (PDB)	www.rcsb.org/pdb
Nucleic Acids Database (NDB)	ndbserver.rutgers/edu/
HIV Protease Database	www.ncifcrf.gov/CRYS/HIVdb/NEW_DATABASE
ReLiBase	www2.ebi.ac.uk:8081/home.html
PDBsum	www.biochem.ucl.ac.uk/bsm/pdbsum
CATH	www.biochem.ucl.ac.uk/bsm/cath
SCOP	scop.mrc-lmb.cam.ac.uk/scop
FSSP	www2.embl-ebi.ac.uk/dali/fssp
Nucleotide sequences	
GenBank	www.ncbi.nlm.nih.gov/Genbank
EMBL	www.ebi.ac.uk/embl
DDBJ	www.ddbj.nig.ac.jp/
Genome sequences	
Entrez genomes	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
GeneCensus	bioinfo.mbb.yale.edu/genome
COGs	www.ncbi.nlm.nih.gov/COG
Pathway	
KEGG	http://www.kegg.com/
DIP	http://dip.doe-mbi/ucla.edu/
BIND	http://www.blueprint.org/bind/bind.php
Integrated databases	
InterPro	www.ebi.ac.uk/interpro
Sequence retrieval system (SRS)	www.expasy.ch/srs5
Entrez	www.ncbi.nlm.nih.gov/Entrez

터들은 24시간 교류되고 있다.[9,19,20] 생물정보로 이용되는 데이터베이스는 문헌 데이터베이스, 유전체 서열, 단백질 서열, 구조 데이터베이스로 분류되며,[3,9,18] 유전체 데이터베이스의 경우 human, mouse, rat, zebrafish, fungus 등이 있고, 유전체의 특징 표적 부위를 저장해 놓은 ESTs, STSs, EPD, REPBASE, HTG, GSS 등이 있다.[9,21] 단백질 서열의 관련 데이터베이스는 단백질 서열 정보 이외에 단백질의 기능과 구조, 각 도메인에 관한 정보, post-translational modification, variants에 대한 정보 등 관련된 모든 정보들을 잘 가공하여 구축한 SWISS-PROT으로 2003년 12월 기준으로 141,681개의 서열들이 저장되어 있고, EMBL의 유전자 서열들을 해독시켜 구축한 SWISS-PROT의 1차 데이터베이스로 2003년 12월 현재 1,078,339개의 서열들에 관한 정보들을 수록하고 있는 TrEMBL, 구조 데이터베이스인 PDB를 비롯하여 PIR, PRF, 도메인, 모티브 및 protein 패밀리의 PROSITE, Pfam, ProDom 등으로 구성된다. 그 외 NCBI의 accession number를 가지는 DNA, RNA, Protein 데이터를 제공하는 RefSeq 데이터베이스와 DNA 서열 데이터베이스를 단백질

의 데이터로 변환한 SPTrEMBL, GenBank에서 유전자의 coding 부위를 변환한 GenPept, 일본 교토 화학연구소 GenomeNet의 pathway 데이터베이스인 KEGG, 인간 유전병과 유전체의 돌연변이에 대한 데이터베이스로 OMIM, HGMD 등이 있다.[23] 또한 단백질 구조 분류에 대한 데이터베이스는 구조와 기능의 상동성을 가지는 단백질들을 공통된 조상으로부터 그룹화하여 계층 구조를 분류시키면서, 서열 유사성 검색에 기초를 두고 구조비교와 수작업 조사에 의해 단백질 구조를 군집화 하는 SCOP과 반 자동 구조 분류 시스템으로 단백질 구조 비교 프로그램인 SSAP를 구조 분류에 이용하는 CATH, 그리고 프로그램 DALI의 구조 비교 시스템을 이용하여 완전 자동화된 분류 시스템으로 되어진 FSSP 등이 있다.[9,22,23]

데이터베이스를 이용하는 대표적인 분석 도구에는 1981년 Smith & Waterman 알고리즘을 적용한 FASTA가 1988년 Pearson과 Lipman에 의해 등장했으며, 1990년 Altschul 등에 의한 BLAST, 1994년의 Thompson 등에 의해 개발된 CLUSTALW가 대표적인 분석도구로 사용되고 있다.[18,24,25] 기본적으로 바이오인포매틱스의 분석 도구는 서열 데이터간의 유사성, 상동성 및 계통관계를 알아보거나 특이성을 발견하기 위해 사용하게 되며, 수집 저장된 데이터에 주석을 달거나 단백질의 상호작용과 같이 데이터간의 연관성과 in silico에서의 데이터 조절 및 유전형에서 표현형을 유추하는 등 많은 부분에 적용되고 있다.[10,16,26,27,28,29] 그 중 서열 유사성 검색도구로 대표되는 도구는 1990년대 초에 개발되어 최근까지 일반적으로 사용되는 BLAST이다. FASTA와 BLAST는 서열 유사성 도구라는 점에서 유사하게 취급되고 있으나, FASTA는 검출감도 면에서, BLAST는 처리속도 부분에서 우수하다고 할 수 있다. 영국 Sanger 연구소의 hashing 알고리즘을 사용하는 SSAHA는 DNA와 Protein 데이터베이스의 빠른 검색을 지원하는 유사성 서열 검색 프로그램으로 염색체상의 동일 서열 부위 등을 찾는데 이용되고 빠른 특징을 가지고 있다.[21,24,25] 현재 가장 많이 사용되는 도구는 BLAST이고, 아미노산 부분에서는 PSI-BLAST와 PHI-BLAST가 사용되며, 염기서열의 분석도구로는 BLAST2.0이 이용되고 있다. 아미노산과 DNA 서열의 데이터베이스를 기반으로 하여 검색하는 BLAST의 구성은 BLASTN, BLASTP, BLASTX, TBLASTN, TBLASTX로 구성되며, 아미노산 서열과 염기서열 및 변환서열(6 frame translated nucleotide)을 적용할 수 있다.[18,25] 서열 정열 프로그램에 속하는 CLUSTALW는 다양한 소프트웨어 패키지에 적용되고 있다. 또한 그 알고리즘이 응용되어 시각화 요소를 극대화한 새로운 브랜드의 제품들에 사용되고 있으며, 다수 서열의 정열 특성상 계통적인 것으로 많은 부분이 나타나므로 계통도를 나타내는 도구들과 연동되어 분석결과를 보여주게 된다. 이와 더불어 서열의 특정 부분만을 검색하여 분석하도록 도와주는 NCBI의 ORF Finder, VecScreen 등과 같이 특이부분에 적용되는 수많은 무료 프로그램들과 세계 바이오인포매틱스 전문회사들의 소

Table II–*Bioinformatics Related Software Products*

제작사	제품명	세부사항
InforMax	Vector NTI Advance Vector Xpression Vector PathBlazer LabShare GenoMax	DNA/Protein 서열 분석 패키지, 데이터베이스 관리, 서열단편 조합, 물리/화학적 특성 분석, 자동/수동 annotation기능, 구조분석 및 시각화, 유전자 발현패턴분석, pathway 분석 및 단백질 구조 예측, EST 지도 완성, 세포내 단백질 역학분석, 통합분석 환경
BioMax	Pedant-Pro HarvESTer BioRS BioXM Human/Pedant Genome DB	EST clustering과 assembly 시스템, EST과 유전체의 자동/수동 annotation 시스템, 데이터정보의 retrieval 시스템, 생물정보분석 tool box 제공, Chip 분석, 다양한 annotated 유전체 DB 제공, human annotation DB제공
Accelrys	Accord, Catalyst Cerius^2 CNX, FELIX, GCG, Insight II, Matcrials Studio, MedChem. Explorer, Quanta, RS^3 TOPKAT, TSAR, WebLab	분자 구조 모델링, 조립화학 및 신약설계, 시뮬레이션 패키지 및 부석 주제별 다양한 application, 화학물질 데이터베이스, 관리 프로그램, 화학물질의 구조화 및 검색, 서열분석 프로그램, 미들웨어 기술, DNA, RNA, Protein 분석
	Pro-Chart	Protein-Protein interaction DB
Geospiza	Finch-Suite, i-Finch	LIMS 체제의 contig assembly
Media Cybernetics	Array-Pro, Gel-Pro	Chip 이미지 분석, gel 이미지 분석

프트웨어 프로그램 들은 서열분석, 대량 유전자 발현 분석, 구조분석, 대사경로 분석 등에 적용되어 통계적이고, 전산적인 분석적용이 어려운 생물학자들에게 획기적인 도움이 되고 있다. 이러한 생물학자들을 위해 상용화된 데이터 분석도구의 가장 첫 번째 조건이 웹 기반의 친숙한 환경의 인터페이스와 분석과정을 쉽고 빠르게 적용시켜 원하는 데이터를 만들 수 있도록 하는 통합의 기능이라고 할 수 있다. 현재 출시 된 상용프로그램(Table II) 중 가장 통합기능이 우수하고 다양한 기능을 갖추고 있는 프로그램은 미국 InforMax사의 'Vector NTI Suite'로 이를 포함하여 vector family에 속하는 Xpression과 PathBlazer가 있다.[30,31] 이것은 대량 유전자 발현 분석, 서열분석, 단백질 상호작용 및 대사경로 분석을 포함하여 통합화를 최적화 하였고, 각각 다른 성질의 데이터를 데이터베이스화하여 처리한다는 점에서 다른 프로그램과 차별화 된다고 할 수 있다. 5개의 application molecule로 수행되는 Vector NTI Suite는 데이터 분석과 관리도구를 포함하며, sequence creating, mapping, analysis, annotation 등을 수행하는 Vector NTI 모듈과 DNA 염기서열의 assembly를 담당하는 ContigExpress 모듈을 사용하여 특이적인 데이터의 표출과 형식에 대해 다양한 데이터 파일 적용(GeneBank, FASTA, ABI, SCF, ALF, EMBL, Text, ESD, Phred. Phrap ACE)이 가능하다. 또한 GenomeBench 모듈은 desktop 상에서 인간 유전체 프로젝트의 개별적인 유전체 데이터를 UCSC (University of California Santa Cruz), Ensembl, TIGR, WORMBASE 등의 DSA(distributed annotation system) server로부터 annotated features를 다운로드 받아 local에 있는 자신의 데이터와 연동하여 Spidey나 SIM4의 분석프로그램을 사용하여 reference genomic backbone에서 위치를 확인할 수 있도록 하고, genomic backbone을 local 데이터베이스에 저장하여 정렬 할 수 있다. 이는 유전체 데이터를 연구하는데 있어 매우 중요한 기능으로 mRNAs, ESTs, STSs와 같은 정보를 GenBank에 존재하는 다른 특징까지 포함하여 활용할 수 있도록 한다. 특징적인 염기서열 정보의 물리화학적인 특성을 고려한 Bio-annotator 모듈과 복잡한 염기서열 정보를 배열하는 주요기능으로 서열간의 유사성을 나타내어 계통도를 guide tree로 보여주는 AlignX로 구성된다. Explorer라고 하는 데이터베이스는 local에 자동으로 구축되는 관리도구 개념의 데이터베이스로서, 이러한 데이터베이스를 기반으로 저장된 데이터를 통합적인 시스템 형태를 갖추고 있기 때문에 생물학자들에게 쉽고 편리한 인터페이스를 제공하여 통합분석을 실현할 수 있는 최적의 in silico 환경을 제공하게 된다. 최근 데이터분석의 흐름은 데이터의 주석을 달아 데이터의 가치를 부여하는 방향으로 나아가고 있으며, 독일 Biomax사의 Pedant-Pro가 유전체의 auto/manual-annotation 기능을 최상으로 공급함으로써 이러한 분석 흐름의 선두를 지켜가고 있다.[32] 여기에 사용되는 computational method에는 22개의 프로그램이 사용되고, 12개의 데이터베이스가 활용된다. ALOM2는 개별 염기서열의 transmembrane region을 검출하고, gap이 없는 blocks의 데이터베이스를 활용하여 protein block의 유사성 검색을 하는 BLIMPS, 단백질 코일의 score 분포를 이용하여 코일 형성부위를 예측하는 프로그램인 COILS, 유전체 서열에서 유전자를 찾는 annotation 프로그램인 DDS와 DPS, 유전자를 예측하기 위한 유사성 비교와 HMMs (hidden markov models)을 사용하는 Fgenesh++, 생물학 데이터베이스 염기서열 파일 변화 프로그램에 속하는 fmtseq, 유전자 예측 프로그램의 Genscan, HMMs을 만들기 위해 multiple aequence alignment 데이터를 사용하는 HMMER, 유사성 서열 검색의 NCBI-BLAST, 박테리아 유전체와 대량 유전체 단편의 유전자 예측 시스템인 ORPHEUS, 개별적인 염기서열과 완성된 유전체 데이터로부터 기능적, 구조적 특징을 제공하는 PEDANT, Perl script의 pI, 높은 정확성을 가진 2차 구조 예측 시스템의 PREDATOR, PROSITE 데이터베이스의 단백질 서열을 결정하기 위한 서열 검색 시스템인 ProSesrch, 단백질의 지역 위치를 예측하는데 사용되는 PSORT, 단백질 서열의 통계적인 분석을 시도하는 SAPS, 단백질 서열의 low-complexity 부위를 검출하는 SEG, signal peptide 예측 프로그램인 SignalP, HMM을 기본으로 하여 membrane-spanning 단백질에서 alpha helices의 위치와 방향을 예측하는 TMHMM, 유전체의 DNA 혹은 RNA의 tRNA 유전자 규명을 위한 프로그램인 tRNAscan-SE가 auto annotation 분석에 사용되고 있다. 또한 이들 프로그램이 사용하는 데이터베이스는 protein block 데이터베이스에 Blocks, 단백질의 계통적인 연관 관계를 집단적으로 구성하는 COGs, 유럽의 1차 염기서열 리소스로 사용되는 EMBL, 단백질 구조데이터베이스에 PDB, protein domain 데이터베이스의 Pfam, 단백질 정보제공과 국제적 단백질 서열의 PIR-PSD, 스위스 생물정보학 기관에서 구축한 단백질 families와 domain의 PROSITE, 단백질 기능에 대한 분류를 정리한 functional category 데이터베이스, atomic coordinates로부터 2차 구조 성분을 인지하는 STRIDE, 단백질의 구조적, 진화적 상관관계를 고려한 구조분류 데이터베이스인 SCOP, 스위스의 단백질서열 데이터베이스 SWISS-PROT, EMBL에 의해 유지되는 데이터베이스로 모든 암호화 서열의 변환 서열을 저장해 놓은 TrEMBL 데이터베이스가 있다.

바이오인포매틱스 데이터베이스를 활용한 서비스는 크게

네 가지로 분류 할 수 있다. 바이오 데이터베이스의 통합과 데이터 마이닝, 유전자 기능 예측 등의 데이터 분석 그리고 시뮬레이션을 활용한 그래픽 인터페이스의 시각화이다.[18,33] 이러한 데이터들의 최대한의 활용은 바이오인포매틱스의 발전이 될 수 있고, 그러기 위해 데이터베이스를 기반으로 하는 프로그램의 성능과 활용하는 연구원들의 수준은 중요한 문제가 된다. 따라서, 복합 학문성향의 바이오인포매틱스는 여러 학문의 범주를 넘나드는 인재 양성이 요구되고 있다.

참여기업

바이오인포매틱스 분야에서 활약하고 있는 소프트웨어 기업으로는 InforMax, Geospiza, BioMax, LION Bioscience, Accelrys, Media Cybernetics 등이 있고, 생물정보 서비스 기업으로는 Incyte Genomics, Celera Genomics, CuraGen 등을 비롯해 무수한 기업들이 있다. 이와 더불어 정보기술 인프라 기업으로 IBM, Hewlett Packard, Hitachi 등이 있다.[8,32,34] 상세히 국내의 생물정보 소프트웨어 및 서비스를 제공하고 있는 업체는 두 부류로 나뉠 수 있다. 외국의 본사를 바탕으로 시스템 통합을 구축하고 해당 소프트웨어의 판매 및 컨설팅을 지원하는 업체들과 소프트웨어의 자체 개발에 힘을 쏟고 있는 업체들로 나뉜다.

바이오인포매틱스의 향후 전망

바이오인포매틱스는 컴퓨터를 이용하여 대규모의 생물정보를 효율적으로 검색, 처리, 저장 및 분석할 수 있고, 기능 유전체학과 프로테옴 연구 및 개체간의 대사경로에 이르기까지 생산된 원조 데이터를 활용하여 또 다른 의미의 정보를 만들어 낼 수 있는 바이오 산업 발전을 위한 가장 핵심적인 분야라고 할 수 있다. 앞으로는 생물학 연구의 50 – 70%가 바이오인포매틱스에 의해 이루어질 것이라는 견해가 있으며 이는 생물산업 발전을 위한 바이오인포매틱스 분야에 대한 준비가 필요하다는 것을 말한다. 이에 앞선 선진국들의 바이오 혁명의 움직임은 20세기 인간 유전체 프로젝트의 시작과 더불어 데이터베이스 접속 방법인 전자우편서버, FTP(File Transfer Protocol), TELNET 서버 등으로부터 WWW(world wide web)을 통한 국제데이터베이스들 간의 상호접속과 검색을 용이하게 하는 강력한 네트워크 인프라를 통해 시작되었다고 볼 수 있다.[8,35] 이로 인해 지금까지 생성된 수 많은 데이터 중 공개된 데이터와 인간유전체프로젝트를 비롯하여 세계 각 국의 허용된 연구프로젝트의 진행결과를 쉽게 확인할 수 있고, 또한 외부 데이터와 자신의 데이터를 비교 분석하는 것이 당연한 실험 분석과정의 흐름으로 자리한지 오래라고 할 수 있다. 현재 바이오인포매틱스의 패러다임은 생산된 데이터를 거슬러 올라 컴퓨터와 분석도구들을 활용하여 유전자 서열의 감추어진 발현성 및 표현형을 찾아가는 것으로 연구흐름의 방식이 진행되고 있다.[15,36] 그러므로 기존의 데이터를 생명현상과 가까운 의미 있는 데이터로 만들어 가는 작업은 복잡한 단계의 현상을 하나씩 거꾸로 풀어가는 효과를 지닌 동시에 수많은 유전자를 함께 연구할 수 있다는 점에서 이를 바이오인포매틱스의 기술적인 동향으로 볼 수 있다. 또한 이러한 역할은 기존 학문의 보조적인 개념을 넘어 세계 IT 기업들을 생명과학 시장으로 뛰어들게 만들어 산업적으로 그 가치가 매우 높은 생명체의 정보기술 분야에 눈을 뜨게 만들었다. 이러한 정보기술이 바이오산업에 응용되어 얻어지게 될 결과는 생명과학이라는 분야가 정보기술의 혜택을 가장 늦게 도입하였음에도 불구하고 유전정보를 알고자 하는 인간의 욕구와 생명의 신비에 한 발짝 다가서는 중요한 단서가 된다는 점에서 그 파급효과가 대단히 클 것으로 기대된다. 그뿐 아니라 바이오인포매틱스로 인한 생물정보의 급격한 성장은 생물체의 유전 정보를 조작 할 수 있게 되는 가능성이 점점 높아진다는 차원에서 반향적인 사회적 파장 또한 가져올 수 있을 것으로 판단된다. 이를테면 개인 유전자 정보의 프라이버시와 생물체 정보에 기인한 유전자 맞춤 치료들은 21세기 우리가 앞으로 끝없이 그 수준을 조율하며, 생물정보기술의 혜택의 폭을 어디까지 허용할 것이냐 하는 윤리적인 문제까지 생각할 수 있을 것이다. 그러나 여기서 분명한 것은 바이오 산업 전체와 정보기술의 문제인 바이오인포매틱스의 성장은 매우 밀접한 연관성이 있다는 것을 알 수 있으며, 앞으로 넘쳐나는 생명과학 데이터에 정보기술을 도입한 이들의 경제적, 사회적 잠재력은 21세기 생물산업을 주도적으로 이끌어 가기에 충분하다는 것이다. 따라서, 바이오 정보기술을 이용한 바이오인포매틱스의 성장과 응용부분이 바이오 산업혁명의 주축이 될 것으로 생각되며, 생명 정보의 정복시대에 대비하여 바이오인포매틱스의 흐름을 주도해 나아갈 필요성이 요구된다.

감사의 글

본 논문은 보건복지부 보건의료기술진흥사업(01-PJ1-Pg1-01CH06-0003)과 BK21의 지원에 의해 이루어진 것임.

문 헌

1) The British Council Website (http://www.britishcouncil.org.in/science), Celebrating 50 years since the discovery of DNA structure in the UK, British Council (2003).

2) Erwin Chargaff Papers, American Philosophical Society (1993).
3) N.M. Luscombe, D. Greenbaum and M. Gerstein, What is bioinformatics? An introduction and overview, *IMIA Yearbook*, (http://bioinfo. mbb.yale. edu/~nick/ bioinformatics/) (2001).
4) National Human Genome Research Institute website [http:// www.genome.gov].
5) B. Rybak, Bio-Informatics and Bio-Process in the Physiology of Communication, *Biosciences Comm.*, **4**, 158-159 (1978).
6) D. Story, An Understandable Definition of Bioinformatics., News & Articles, OReilly Network Weblogs, Feb. **5**, (2003).
7) The OReilly website [http://www.oreilly.com/news/bioinformatics_0401.html].
8) S.K. Kim, The combination of biotechnology and technology information, Science and Technology Policy Institute, (2002).
9) The National Center for Biotechnology Information website [http://www. ncbi.nlm.nih.gov].
10) I. Iliopoulos, S. Tsoka, M.A. Andrade, P. Janssen, B. Audit, A. Tramontano, A. Valencia, C. Leroy, C. Sander and CA Ouzounis, Genome sequences and great expectations, *Genome Biol.*, **2**, 1-3 (2001).
11) J.H. Park and K.S. Han, Challenges and New Approaches in Genomics and Bioinformatics. *Genomics & Informatics*, **1**, 1-6 (2003).
12) E.J. Ko, Connection of biotechnology and information technology, *LG Economic Week*, **652**, 22-27 (2001).
13) B.M. Kwon and J.H. Yu, Chemical genetics and chemical proteomics, *Biochemistry News*, **20**, 12 (2000).
14) C.K. Huh, Role of bioinformatics in functional genomicse, *Biochemistry News*, **20**, 1 (2000).
15) E.J. Ko, Rising the usage of biotech, *LG Economic Week*, **678**, 38-42 (2002).
16) G. Cagney, S. Amiri, T. Premawaradena, M. Lindo and A. Emili, In silico proteome analysis to facilitate proteomics experiments using mass spectrometry, *Proteome Sci.*, **1**, 5 (2003).
17) R. Overbeek, Genomics: What is realistically achievable?, *Genome Biol.*, **1**, comment2002.1-2002.3 (2000).
18) E.Y. Kim, Bioinformatics, Korea Institute of Science and Technology Information (2003).
19) European Bioinformatics Institution website [http://www.ebi. ac.uk/].
20) DNA Data Bank of Japan website [http://www.ddbj.nig.ac.jp/].
21) The Ensembl Genome Browser website [http://www.ensembl. org/].
22) K.H. Rhyu, Bioinformatics and database research direction in post-genome age, *Korea Institute of Science and Technology Information*, **12**, 108-117 (2003).
23) A. Karwath and R.D. King, Homology induction: The use of machine learning to improve sequence similarity searches, *BMC Bioinformatics*, **3**, 11-24 (2002).
24) S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403-410 (1990).
25) A. Pertsemlidis and J.W. Fondon, Having a BLAST with bioinformatics (and avoiding BLASTphemy), *Genome Biol.*, **2**, reviews2002.1-2002.10 (2001).
26) M. Lexa, J. Horak and B. Brzobohaty, Virtual PCR, *Bioinformatics*, **17**, 192-193 (2001).
27) P.E. Hodges, P.M. Carrico, J.D. Hogan, K.E. O'Neill, J.J. Owen, M. Mangan, B.P. Davis, J.E. Brooks and J.I. Garrels, Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from Incyte Genomics, *Nucleic Acids Res.*, **30**, 137-141 (2002).
28) J.H. Graber, C.R. Cantor, S.C. Mohr and T.F. Smith, In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species, *Proc. Natl. Acad. Sci.*, **96**, 14055-14060 (1999).
29) J.S. Edwards, R.U. Ibarra and B.O. Palsson, In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data, *Nat. Biotechnol.*, **19**, 125-130 (2001).
30) The LG economic research institute website [http://www.lgeri. com/].
31) The InforMax website [http://www.informaxinc.com/content. cfm?pageid=1].
32) The Biomax Informatics AG website [http://www.biomax.de/].
33) The Korea institute of science and technology information website [http://www.kisti.re.kr/kisti/knowledge/knowledge_main. jsp?sub=3].
34) The Bitzkorea website [http://www.bitzkorea.com].
35) D. Rocco and T. Critchlow, Automatic discovery and classification of bioinformatics web sources. *Bioinformatics*, **19**, 1927-1933 (2003).
36) C.A. Semple, M.S. Taylor and S. Ballereau, The mesogenomic era. *Genome Biol.*, **2**, 4015 (2001).