특 집

Bio-Informatics 개론 및 국내외 동향

이 식
한국과학기술정보연구원 바이오인포매틱스센터

I. 서 론

20세기말에 시작되어 21세기초에 완료된 인간 유전체사업(Human Genome Project, 이하 HGP)은 인간이 이룬 과학적 업적 중에 손꼽히는 성과로 기억될 것이다. 사람의 DNA 염기서열 전체를 결정하는 HGP가 가능하게 된 것은 분자생물학의 발전, 자동화된 장비의 발명, 그리고 고속 컴퓨터와 효율적인 알고리즘의 등장으로 고속처리(high-throughput) 연구가 가능해졌기 때문이다.

유전체(genome)란 세포의 핵에 존재하는 염색체의 절반, 즉 반수체(haploid) 염색체 내에 들어있는 유전자의 집합을 말한다. HGP 등 유전체사업의 결과로 얻어지는 생물학적인 정보를 처리하여 유용한 정보를 끌어내는 것이 바이오인포매틱스이다. 바이오인포매틱스는 생물학 및 의학분야 연구에 있어서는 필수적인 분야로 자리잡았고 점차 주변학문과 융합되면서 범위가 넓어지고 있다. 때문에 바이오인포매틱스는 넓은 의미로는 컴퓨터를 이용하여 생명과학분야를 연구·활용하는 모든 분야를 아우르는 학문이라 할 수 있다.

선진국보다는 좀 늦은 감이 있지만, 국내에서도 최근 들어 바이오인포매틱스가 유행처럼 번지고 있다. 바이오인포매틱스가 이처럼 주목받는 것은 유전정보의 효율적인 활용여부에 따라 인간의 삶에 미칠 파급효과가 엄청나게 크기 때문이다. 과히 새로운 산업혁명 내지는 과학혁명이라 할 수 있을 것이다.

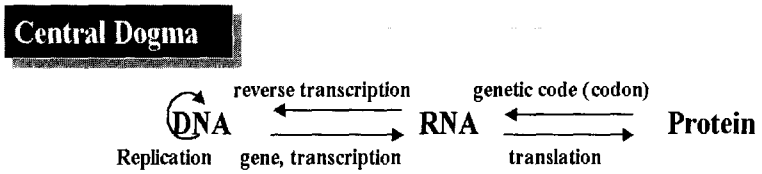
본 리뷰에서는 유전정보의 흐름과 생물학 데이터, 바이오인포매틱스의 정의, HGP의 소개, 바이오인포매틱스의 주요 연구분야, 마지막으로 그리드 기술의 활용 등을 소개한다. 이번 특집호의 다른 리뷰에서 다루고 있는 데이터 마이닝, DNA Microarry 분석, 인공지능, 데이터베이스 등에 대해서는 간략하게만 언급하였다.

II. 유전정보의 흐름 및 생물학 데이타

생명현상이 유지되기 위해서는 정해진 유전정보가 주변환경에 반응하여 적시에 발현되어야 한다. 유전정보를 담고 있는 DNA는 복제를 위해 스스로 주형이 되고, DNA로부터 RNA가 전사되며, RNA는 단백질로 번역된다. 즉, 유전정보는 DNA에서 RNA로, RNA에서 단백질로 전달된다. 중심원리(central dogma)로 불리는 이 명제는 생명정보의 흐름을 간단하게 도식화한 것으로, 생명을 지속시키기 위한 모든 마스터플랜이 여기에 담겨있다.

DNA는 아데닌(A), 티민(T), 구아닌(G), 사이토신(C)의 네 가지 염기로 이루어져 있으며, RNA에는 티민(T) 대신 우라실(U)이 존재한다. 단백질은 20가지의 아미노산으로 이루어져 있다. 즉, DNA, RNA, 단백질은 거대분자이긴 하지만 종류가 명확하게 규정된 구성물질로 이루어진 선 모양의 사슬이므로 기호의 서열로 표시할 수 있다.

1953년 케임브리지 대학교의 왓슨과 클릭에 의



〈그림 1〉 분자생물학의 중심원리

해 DNA가 수소결합을 통해 이중나선 구조를 이루고 있음이 밝혀지면서 분자생물학의 시대가 시작된다. 특히, DNA를 원하는 자리에서 마음대로 자를 수 있는 제한효소(restriction enzyme)의 발견, DNA의 양을 증폭시킬 수 있는 PCR (polymerase chain reaction)의 개발, 그리고 자동화된 염기서열판독(sequencing) 기술의 발달로 전체 유전체 연구는 더욱 가속도를 받고 있다.

고등생물의 경우 전체 유전체 서열 중의 극히 일부분만이 단백질을 암호화하는 영역(coding region)이다. DNA 서열 중 실제로 단백질 합성 정보를 담고 있는 엑손(exon)을 1차적으로 선발해 내는 과정이 필요한 이유이다. 이 때문에 엑손과 인트론(intron)이 불규칙적으로 산재해 있는 전체 유전체 서열 중에서 의미 있는 부분만 전사된 mRNA가 주목받고 있다.

RNA은 DNA에 비해서 불안정하기 때문에 바이러스에서 추출한 역전사효소(reverse transcriptase)를 이용하여 mRNA에 상보적인 cDNA를 합성하여 이의 서열을 결정하게 된다. 특히, 양쪽 끝에서 대략 500개 정도 읽은 서열을 EST(expressed sequence tag)라 한다. EST는 mRNA를 이용하여 만들어진 것이기 때문에 전체 DNA 서열 중 엑손 부분만 포함된 정보를 가지고 있다. 때문에 염색체 DNA에서 유전자의 정확한 위치를 찾거나 질병의 연구 등에 유용하게 쓰인다.

유전체에 대한 기본적인 개념과 구조, 생물학적 특성 등을 이해하고 있다면 살아있는 생물체 내에서 일어나는 모든 현상은 수많은 정보의 흐름임을 이해할 수 있다. 정보의 주체는 당연히 유전자이다. 유전자는 유전체의 한 부분으로 실제로 특정 기능을 담은 정보를 가지고 있는 기본적인 단위로 이해할 수 있다. 이러한 정보에 접근하기 위해서 가장 먼저 해야 할 일이 1차원적인 정보인 서열데이타를 얻는 일이다. 이러한 이유 때문에 HGP 같은 서열데이타 수집전쟁이 벌어진 것이다. 하지만 유전자 전쟁의 근본적인 목적은 유전자가 언제, 어디서, 어떻게 발현되는지를 이해하는 것이다.

HGP의 영향으로 초기에는 DNA나 단백질의 서열정보 분석에만 치우친 경향이 많았지만, 포스트-지놈(post-genome) 시대에 접어들면서 유전자의 기능, 단백질의 3차 구조, 기능 및 네트워크 등을 정확하게 규명하는 일의 중요성이 부각되고 있다. 이것은 곧 질병에 대한 치료약 개발 등과도 직결된다. 참고로 2003년 10월 기준의 Genome Online Database에 의하면 전세계적으로 이미 서열판독이 끝난 유전체가 162종, 현재 진행중인 프로젝트가 672개(원핵생물 410, 진핵생물 262)에 이른다(http://igweb.integratedgenomics.com/GOLD/).

III. 바이오인포매틱스의 정의와 중요 데이터베이스

1. 등장 배경 및 역사

생물학분야는 다른 분야와 마찬가지로 주변 학문의 영향을 받으며 단계적으로 발전해 왔다. DNA의 이중나선구조가 발견된 1953년 이래로 분자생물학과 유전학, 생화학을 중심으로 생물학은 20세기 후반에 비약적인 발전을 이루었고, 이에 따라 생성되는 생물학 데이터의 양이 증가하기 시작하였다.

바이오인포매틱스란 용어가 본격적으로 주목을 받기 시작한 것은 1990년이지만, '바이오인포매틱스'란 용어가 처음 등장한 것은 1960년대이다. 생물학적 데이터를 컴퓨터로 분석한 연구결과가 발표되었고, '바이오인포매틱스'란 용어가 처음 사용된 것도 1960년대이다. 그러나, 1990년대 들어 본격적으로 바이오인포매틱스란 용어가 사용되기 전까지는 전산생물학(computational biology)이란 용어가 더 많이 사용되었다.

초기에는 의학관련 문헌정보, DNA와 단백질 1차 서열 정보, 단백질의 3차 구조 정보 등을 수집하여 비교적 간단한 데이터베이스를 구축하는데 그쳤다. 그러나 서열분석기를 비롯한 각종 자동화된 장비와 컴퓨터의 발달, 그리고 HGP가 진행됨에 따라 생물학 데이터의 생성속도는 더욱 가속화되어 생물학 정보의 양은 기하급수적으로 증가하고 있다. 2003년 2월 기준으로 GenBank에만 약 2232만건의 데이터(총 길이는 285억개)가 수집되어 있다(www.ncbi.nlm.nih.gov/Genbank/genbankstatts. html). 1995년까지는 데이터가 2배로 늘어나는 데 걸리는 기간이 20개월이었으나, 그 이후로 14개월로 짧아졌으며, 앞으로는 더욱 더 짧아질 것으로 전망된다.

DNA와 단백질에서의 서열 정보는 물론이고 이들의 3차원 구조 정보, DNA Microarray 정보, 단백질의 상호작용 네트워크, 문헌 정보 등에서 쏟아져 나오는 데이터 역시 기하급수적으로 증가하고 있다. 전세계에서 기하급수적으로 쏟아져 나오는 방대한 데이터를 수집하고 효율적으로 관리하며 연구·분석에 활용하기 위해서 대량의 데이터를 관리하고 분석할 수 있는 환경의 필요성이 제기되었다. 이에 따라 미국의 NCBI(National Center for Biotechnology Information), 유럽의 EBI(European Bioinformatics Institute), 일본의 CIB(Center for Information Biotechnology)와 같은 정부지원의 전담기관이 설립되었다. 이들 대표적인 기관에서는 DNA의 염기서열, 단백질의 아미노산 서열, 단백질의 2차 및 3차 구조, 데이터 분석을 위한 소프트웨어, 각종 문헌정보 등을 인터넷을 통해 제공하고 있으며, 데이터의 정확성 및 신속성을 위해 상호간에 데이터베이스를 참조하는 협약이 맺어져 있다. 국내의 경우는 한국생명공학연구원과 한국과학기술정보연구원이 협력하여 국가유전체 정보센터(NGIC, www.ngic.re.kr)를 운영하고 있다.

컴퓨터 하드웨어 및 알고리즘의 발전, 데이터베이스, 전세계적인 인터넷의 보급에 따른 정보의 공유화와 다양화가 가능해지면서 컴퓨터를 활용한 데이터 마이닝이 가능하게 되었고 매우 복잡한 생명현상을 다양하게 분석할 수 있는 도구들이 계속 개발되고 있다. 생물학 연구자 사이에서도 바이오인포매틱스를 생물학의 새로운 분야로 인식하게 되면서 바이오인포매틱스는 급속하

〈표 1〉 대표적인 데이터 관리 기관

국가	대표기관	연혁 및 주요 사업
미국	NCBI	· 1988년 국립보건원(NIH) 산하에 설립 · GenBank를 관리하고 있으며, 서열 데이터분석을 위한 소프트웨어 도구의 개발 및 검색 서비스 실시 · 의학 및 생물학 관련 문헌정보 데이터베이스(PubMed) 관리 · OMIM, MMDB, UniGene 등의 유명 데이터베이스 관리 · 통합검색 엔진인 Entrez를 통해 서비스
유럽	EBI	· 1992년 EMBL의 산하기관으로 설립 · 유전자 비교, 유전자 예측, 대사경로, 서열-구조 관계, 단백질 3차원 구조, 데이터베이스 링크) 등을 서비스
일본	CIB	· 국립유전학연구소(NIG) 산하기관으로 1995년 설립 · 진화/분류학 관련 서비스에 치중

게 붐을 이루기 시작한다.

바이오인포매틱스는 생물학은 물론이고 약학, 의학, 수학, 통계학, 전산학, 물리학, 화학, 문헌정보학, 공학 등과 밀접한 관계를 맺고 있는 대표적인 학제간 연구분야이다. 이처럼 다양한 분야의 연구자들이 자신들의 방식으로 생물학 연구에 적용하고 있기 때문에 바이오인포매틱스의 정의, 분류, 범위가 사람에 따라 상이하다.

2. 바이오인포매틱스의 정의

생물정보학 또는 생명정보학으로 번역되는 바이오인포매틱스는 생물학을 의미하는 bio와 정보학을 뜻하는 informatics의 합성어이다. 정보학은 데이터의 수집·조직화·해석을 위해 데이터베이스 또는 다른 컴퓨팅 툴을 생성·개발·조작하는 것을 말한다. 즉, 바이오인포매틱스는 정보학적인 접근방법을 사용하여 생물학적인 문제를 해결하는 것이라 할 수 있다. 이 때문에 바이오인포매틱스를 IT와 BT의 결합이라고 부르기도 한다.

바이오인포매틱스는 전산생물학(Computational Biology) 또는 Computational Molecular Biology라고도 부르며, 초기에는 아래 〈표 2〉과 같이 컴퓨터를 활용하여 데이터를 관리하고 분석하는 데 주로 사용되었다.

생물학 연구에도 새로운 실험기법이 계속적으로 도입되고 있다. 자동화된 염기서열(DNA Sequencer), DNA 마이크로어레이 같은 BT 기술은 물론이고 직렬 질량분석기와 2D NMR 처럼 화학에서 사용되는 분광학적 방법, 분자 모델링도 활용되고 있다. 이러한 각종 실험장비에서 얻어지는 데이터의 처리와 분석을 위해서 수학, 통계학, 인공지능 기법은 물론이고 이미지, 시그널, 패턴 등을 다루는 여러 가지 전산학적 방법이 바이오인포매틱스 연구에 널리 사용되게 되었다. 이에 따라 바이오인포매틱스의 범위와 다루는 데이터가 계속해서 증가하고 있으며, 바이오인포매틱스의 정의도 점차 넓어지고 있다.

〈표 2〉 초기의 바이오인포매틱스의 주요 연구분야

초기의 바이오인포매틱스의 주요 분야
· 데이터베이스의 내용과 기능을 향상시키는 일
· 데이터의 생성, 습득 및 주석을 다는 좀 더 나은 도구(tool)을 개발하는 일
· 총괄적인 기능연구를 위해 데이터베이스와 해석도구를 개발하는 일
· 서열의 유사성과 다양성을 나타내고 분석하는 도구를 개발하는 일
· 다양하게 활용될 수 있는 효율적인 알고리즘을 생성하는 일

3. Human Genome Project

바이오인포매틱스가 주목받고 상업화로 나갈 수 있는 데 가장 크게 기여한 것이 바로 HGP이다. 미국 DOE(Department of Energy)와 NIH에서 공동투자하여 1990년에 시작된 HGP에는 미국은 물론이고 영국, 일본, 호주, 중국 등이 참여한 다국적 컨소시움이 중심이 되어 추진되었다. DOE에서는 1997년에는 오크릿지 국립연구소 등이 참여하는 Joint Genome Institute를 설립되었고, 1998년에는 Cellera Genomics에서도 별도의 프로젝트가 시작되었다.

1999년에 염색체 중 처음으로 22번 염색체의 서열결정을 마쳤다. 2000년 6월 마침내 1차 드래프트가 발표되었고, 이 결과는 2001년 2월에 각각 사이언스와 네이쳐에 발표되었다. 2002년에는 생쥐의 드래프트 서열이 발표되었고, 2003년 4월 마침내 HGP의 종료가 선언되었다. HGP과 관련된 자세한 정보는 www.doegenomes.org.에서 얻을 수 있다.

HGP가 완료됨에 따라 인간의 DNA 서열에 주석을 달아서 연구의 편의성을 주는 사이트도 여럿 존재한다. UCSC의 Human Map Viewer, NCBI의 HGP Genome Browser, EBI의 Ensemble Genome Browser 등을 이용하면 인간 유전체 정보에 쉽게 접근할 수 있다.

HGP를 통해 사람의 유전체에서는 1,433,393개의 SNP(single nucleotide polymorphism)가 발견되었다. 약 2,000 베이스당 한 번 꼴로 일어나는 셈이다. SNP와 관련된 정보는 SNP 콘소시움의 웹사이트(http://snp.cshl.org/)에서

찾아볼 수가 있다. SNP 정보를 잘 활용하면 진화의 연구, 범죄자 혹은 친자확인, 유전자형에 따른 맞춤의약품의 개발 등이 가능해진다.

HGP의 완료에 따라 서열의 해석, 유전체의 발현, 단백질체학, 기능 유전학, 구조유전학 등이 더욱 주목받기 시작하였다.

4. 바이오인포매틱스 분야의 중요 데이터베이스

바이오인포매틱스 분야에서 구축된 사설 및 공공 데이터베이스의 수와 종류는 무척 많아서 일일이 나열하기 힘들 정도이다. 그렇지만 크게 나누면, 서열 데이타, 구조 데이타, Chip 데이터, 문헌정보, 네트워크 데이터 등으로 분류할 수 있다. HGP의 영향으로 현재까지는 유전자에 대한 정보가 압도적으로 많은 실정이지만, 앞으로는 다른 분야의 데이터베이스 역시 빠른 속도로 증가될 것으로 기대되고 있다.

핵산 서열 데이터의 경우 NCBI, EMBL, DDBJ 등에서 제공되는 데이터베이스가 대표적이다. 아미노산 서열의 경우는 SIB(Swiss Institute of Bioinformatics)에서 제공하는 SWISS-PROT과 미국·독일·일본의 연구소들이 공동으로 참여하고 있는 PIR(Protein Information Resource)이 가장 유명하다. SWISS-PROT과 관련된 중요한 데이터베이스에는 ENZYME과 PROSITE가 있다. 특히, PROSISE에는 모티프에 대한 정보가 잘 정리되어 있다. TRANSFAC에는 transcription factor에 관한 정보가 잘 정리되어 있다.

구조 관련 데이터베이스 중 가장 널리 사용되고 있는 것은 PDB와 CCDC이다. PDB(Protein Data Bank)는 1971년 미국의 Brookhaven 국립연구소에서 시작된 것으로 단백질 외에도 핵산과 탄수화물에 대한 3차원 구조 정보를 포함하고 있다. 현재는 Rutgers대학과 샌디에고 슈퍼컴퓨터센터, 미국 표준기술연구소(National Institute of Standards and Technology)에서 협력하여 조직한 RCSB(Research Collaboratory for Structural Bioinformatics)에서 PDB를 관리하고 있다. 2003년 10월 기준으로 약 23, 000개의 분자에 대한 구조 정보가 수집되어 있다. 크기가 작은 분자들에 데이터는 CCDB(Cambridge Crystallographic Data Centre)에서 수집하고 있다. 크기가 작은 단백질 분자의 경우는 PDB와 CCBD 양쪽 데이터베이스에서 모두 찾을 수 있다.

단백질 구조에 따라 분류한 별도의 데이터베이스가 존재한다. 단백질 구조를 분류해 놓은 SCOP(Structural Classification of Proteins)과 CATH(Class/Architectura/Topology/Homology), 서열유사성을 이용하여 단백질의 도메인을 자동으로 분류한 DALI, 구조 정렬을 종합해 놓은 CE 등이다.

모티프 데이터베이스는 서열결정을 통해 지금 막 나온 새로운 유전자가 만들 단백질의 기능을 알 수 있는 좋은 지침이 된다. 이러한 목적을 위해 전문가들의 손으로 가공된 Block, RROSITE, Pfam, PRINTS, COG, TIGERFams 등의 모티프 데이터베이스가 만들어졌다. 모티프 데이터베이스는 규모가 매우 작기 때문에 이곳에서 검색에 실패했더라도 실제 서열이 GenBank에는 포함되어 있는 경우가 많으므로 사용시 주의하여야 한다. 단백질간의 상호작용에 관해서는 BIND와 DIP 데이터베이스가 유용하다.

그 외에도 각각의 연구그룹에서 특별한 단백질에 대한 작성한 데이터베이스가 여럿 존재하고 있으며, 면역학 관련된 IGMT, KABAT, MHCPEP 등의 데이터베이스도 있다. 이처럼 단백질의 종류에 따라 분류되어 있는 소규모의 데이터베이스들은 EBI의 홈페이지 http://www2.ebi.ac.uk/msd/Llinks/family.shtml을 통해 접속이 가능하다.

유전체에 담긴 정보 중 실제로 발현되는 부분이 중요하기 때문에 mRNA에서 얻어진 EST의 서열이 주목받고 있다. EST 서열정보가 많아짐에 따라 GenBank에서 별도로 dbEST란 데이터베이스를 구축하여 운영하고 있다. 2003년 10월 기준으로 약 19,000,000개의 정보가 축적되어 있다.

대사경로에 대한 리소스로는 What Is There

(WIT, wit.mcs.anl.gov/MIT2)와 KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg), EcoCys (http://www.ecocyc.org) 등이 유명하다.

문헌정보의 경우 미국 의학도서관(US National Library of Medicine)에서 구축한 MEDLINE이 가장 유명한데, NCBI의 PubMed를 통해 검색이 가능하다. NCBI는 Entrez란 통합 검색 환경(www.ncbi.nl.nih.gov/Entrez)을 통해 문헌정보, 단백질, 핵산, 구조, 유전체, OMIM(Online Mendelian Inheritance in Man) 등의 정보를 제공한다. 모든 정보들은 하이퍼링크되어 있다.

복잡하고 다양한 생물학 정보를 효율적이면서 추후에 확장이 가능한 데이터베이스로 만드는 것은 바이오인포매틱스에서 매우 중요하다. 전세계 여러 곳에 산재한 데이터베이스간의 상호 연결 역시 중요한 이슈이다. 이 때문에 데이터베이스 관련된 일이 바이오인포매틱스에서 가장 중요한 부분을 차지하고 있다. 일반 연구자들의 경우 중요 데이터베이스에 대한 충분한 정보를 가지고 이를 잘 활용하여야 한다.

Ⅳ. 바이오인포매틱스의 주요 연구분야

바이오인포매틱스에 다루는 정보는 유전자 정보에서부터 단백질 3차 구조 및 기능예측, DNA 마이크로어레이를 활용한 유전자 발현, 단백질체학(proteomics), 생화학 반응 경로에 이르기까지 매우 다양하다. 유전체는 물론이고, transcriptome, proteome, metabolome, physiome 등 생물학 관련된 모든 분야에 바이오인포매틱스가 활용되고 있다. 이외에도 신약설계 같은 분자모델링 분야도 바이오인포매틱스 분야에 포함시키는 추세이며, HGP의 완료로 post-genomics 관련 연구가 급속하게 발전하고 있다. 대량의 데이터를 최대한 빠른 속도로 분석(data mining)해야 하기 때문에 인공지능 등을 활용한 효율적인 알고리즘의 설계와 분산컴퓨팅 기법에 대한 연구가 활발하게 수행되고 있다. 최근 들어서는 확장된 분산컴퓨팅 기법인 그리드 기술도 활발하게 도입되고 있다.

〈표 3〉 바이오인포매틱스의 주요 연구분야

분 야	내 용
서열 분석 (Sequence Analysis)	· 서열정렬 · 기능예측 · 유전자 발견
구조 분석 (Structural Analysis)	· 단백질 3차원 구조 및 기능 예측 · RNA 구조 예측
발현 분석 (Expression Analysis)	· 유전자 발현 분석 · 유전자 클러스터링(clustering)
경로 분석 (Pathway Analysis)	· 대사 경로 · 조절 네트워크 (regulartory networks)

1. 데이터베이스

바이오인포매틱스의 시발점이면서 동시에 가장 기본적인 분야라 할 수 있다. 전세계적으로 수 백가지 이상의 유용한 데이터베이스와 이를 효율적으로 사용할 수 있는 각종 소프트웨어와 웹환경의 사용자 인터페이스가 제공되고 있다.

웹 데이터베이스가 과학자들 사이의 정보공유에 크게 기여하게 되면서, 많은 연구자들이 자신의 연구결과를 데이터베이스화 하여 서비스하는데 관심을 갖고 있다. 바이오인포매틱스 데이터베이스 구축과 관리를 위해 데이타베이스의 스키마 설계, 자료관리, 웹 접속 스크립트 관리, XML, 데이터 표준화 등 다양한 분야의 연구가 진행중이다. 자세한 내용은 본 특집의 다른 원고에서 다루어지므로 여기서는 생략한다.

2. 서열정보의 분석

DNA와 단백질은 각각 4 또는 20가지의 구성물질로 이루어진 선 모양의 사슬이므로 기호의 서열로 표현할 수 있다. 이 서열의 유사한 정도를

비교함으로써 분자의 형태 또는 기능을 예측할 수 있게 된다. 이 때문에 분자구조, 진화, 기능 등을 연구하기 위해서 가장 먼저 수행하는 일이 바로 서열간의 상동성(homology) 검색이다. 상동성은 원래 진화적으로 연관성이 있는 경우에만 사용하는 용어였지만, 최근에는 서열의 유사성을 나타내는 일반적인 의미로 확장되어 사용되고 있다. 서열 분석은 데이터마이닝 기법이 가장 많이 사용되는 분야기도 하다.

1980년대 말, 서열 비교를 할 수 있는 알고리즘과 프로그램이 등장하면서 분자생물학 연구는 크게 바뀌게 되었다. 실험에서 얻은 데이터를 데이터베이스 내의 다른 서열데이타와 고속으로 비교할 수 있도록 해주는 짝정렬(pairwise alignment) 방법은 바이오인포매틱스 연구의 가장 중요한 도구로 자리잡았다. 이후 다중 정렬, 계통발생학적 분석, 모티프 분석, 상동성 모델링 프로그램, 웹 기반의 데이터베이스 검색서비스에 이르기까지 BT 연구자 그룹에서 널리 사용하는 많은 분석 도구들은 짝정렬 알고리즘을 기반으로 하고 있다.

서열정렬을 통해 새로운 서열에 대한 기능을 알아내는 일에서부터 단백질의 3차원구조 예측 및 구성, 유전자의 발현에 이르기까지 다양한 분석이 가능하다. 원리적으로는 간단해 보이지만 실제로 구현하는 것은 쉬운 일이 아니다. 핵산에서는 두 서열의 일치여부에 따라 점수를 부여하면 되지만, 아미노산의 경우 비슷한 성질의 다른 아미노산으로 바뀌는 경우가 많다. 이 때문에 아미노산 서열비교에서는 측정행렬(scoring matrix)을 사용한다. 측정행렬로는 진화관계에 기초하여 만들어진 BLOSUM과 포인트 허용 돌연변이(PAM, Point accepted mutation) 행렬이 널리 사용되고 있다. 서열비교시 아미노산 잔기의 삽입(insertion)이나 제거(deletion)에 대해서는 감점을 부여한다. 서열을 비교할 때 사용되는 측정행렬나 갭에 부여하는 감점 등에 따라 완전히 다른 결과를 주기도 하므로 서열비교시 주의할 필요가 있다.

서열비교방법에는 점 표시(dot-plot), 동적 프로그래밍(dynamic programming), 체험적(heuristic) 방법이 있다.

1) Dot-Plot

점 표시는 서열의 일치여부를 행렬상에 점으로 표시하는 간단한 방식으로, 서열이 많이 일치하는 부분에서는 점이 이어져서 선으로 보인다. 이를 이용하여 서열의 일치 정도, 교차, 역위 등을 알아낼 수 있다. 비교하는 서열의 길이(윈도우 사이즈)와 노이즈 표시 옵션을 조절하여 원하는 그림을 얻을 수 있다.

2) 동적 프로그래밍

동적 프로그래밍은 음성인식같은 전산학적인 문제에 주로 활용되던 알고리즘이다. 원래의 문제를 작은 하위 문제들로 쪼개어 최상의 답을 찾는 방식으로 Needleman과 Wunsch에 의해 처음으로 생물학 연구에 도입되었다. 두 서열의 전체를 정렬하는 글로벌 정렬(global alignment) 알고리즘을 Needleman-Wunsch 알고리즘이라고 부른다. 행렬의 왼쪽 위에서 오른쪽 아래까지 스캔하면서 부분 서열에서 고득점인 정렬로부터 최적 정렬이 만들어진다. 글로벌 정렬은 두 서열이 이미 알려져 있고 전체 길이에 걸쳐서 정렬되어야 한다는 전제를 가지고 있다. 즉, 두 서열은 길이가 비슷하고 기능적으로 유사한 경우에만 적용된다.

가장 빈번하게 사용되는 정렬도구는 로컬 정렬(local alignment)이다. 서열의 로컬 정렬을 수행하는 동적 프로그래밍 알고리즘은 Smith-Waterman 알고리즘이라고 한다. 행렬을 따라 최적의 해를 추적할 때 허용되는 추가적인 선택사항을 제외하고는 Needleman-Wunsch 알고리즘과 유사하다. 로컬정렬에서는 두 서열의 처음부터 끝까지를 정렬하는 해를 찾지 않는다. 서열비교시에 누적된 점수가 음의 값이 나오면 그 정렬을 버리고 새로운 로컬 정렬 부분을 찾는다. 사실 전체서열을 비교하는 일보다는, 아직 알려지지 않은 서열을 찾기 위해서 어떤 서열로 서열 데이터베이스를 탐색하는 일이 더욱 빈번하다.

발생학적으로 관련이 있다면 오랜 옛날에 분화되어 서로 다르게 진화한 경우에도 유전자 서열이나 아미노산 서열에 짧은 상동성 부분이 남아 있는 경우가 많다.

동적 프로그래밍 알고리즘은 정확한 해를 주지만, 계산에 많은 시간이 필요하기 때문에 대용량의 데이터에서 사용하기에는 비현실적이다. 때문에, 일상적인 탐색에서는 체험적(heuristic) 방법을 사용하게 된다.

3) 체험적 방법

체험적 방법은 최적의 정렬을 보장하지는 못하지만, 대단위 탐색을 효율적이면서 실용적으로 수행하기 때문에 인간 유전체 같은 대용량의 데이터베이스 대한 검색을 가능하게 한다. FASTA와 BLAST가 대표적인 체험적 방법으로 현재는 대부분의 연구자들이 BLAST(Basic Local Alignment Search Tool) 계열의 프로그램을 사용하고 있다.

BLAST는 지역적 유사성(local similarity)이 있는 부분을 찾아 서열의 짝짓기를 수행하는 알고리즘이다. 질의서열과 정렬을 이루었을 때 미리 지정된 경계값(threshold) 이상의 점수를 가지는 서열의 목록을 미리 만들고, 이 목록을 서열 데이터베이스에서 찾는 방식이다. 이렇게 찾은 짧은 서열들을 확장결합하는 방식으로 작동한다. BLAST는 현재 NCBI에서 만든 것과 워싱턴 주립대학교에서 만든 두 가지 버전이 존재한다. NCBI BLAST는 다중서열의 프로필(profile)을 비교하는 데 초점을 맞추었고, WU-BLAST는 반복부분이나 갭을 처리하는 방식이 약간 다르다.

4) 다중서열 정렬

서로 관련된 여러 가지 유전자나 단백질 서열을 한꺼번에 비교하면 진화적 관계, 기능적으로 중요한 패턴 등을 알아낼 수 있다. 이 과정을 다중 서열 정렬(multiple sequence alignment)이라 하며, 단백질 서열에 가장 보편적으로 적용된다. 비교하고자 하는 다중서열의 개수가 증가함에 따라 계산시간은 기하급수적으로 늘어나기 때문에 다중서열정렬을 할 때는 세심한 주의가 요구된다. 현재 다중서열법에서 가장 널리 사용되는 프로그램은 ClustalW이다.

다중서열을 통해 계통수(phylogenetic tree)를 작성할 수 있고, 프로필이나 모티프 등을 찾을 수 있다. 모티프(motif)란 분자의 기능을 예측할 수 있는 서열의 패턴이나 구조적인 특징을 말하며, 프로필(profile)은 모티프를 기술하는 정량적 혹은 정성적인 방법이다. 계통수의 작성에는 PHYLIP 프로그램이 많이 사용되고 있으며, 프로필과 모티프의 검색에는 위치 특이 측정행렬(PSSM, postion-specific scoring matrix)이나 은닉 마코프 포델(HMM, hidden Markov-Model) 등이 사용된다.

5) 유전자 탐색

서열이 모두 결정되고 난 후에는 각각의 서열에서 유전자를 찾는 작업을 수행해야 한다. 유전자를 찾아주는 응용 프로그램에는 GeneMark, GeneMark.hmm, GenScan 등이 있다. TIGR에서는 각각 원핵과 진핵생물에 적용할 수 있는 Glimmer와 GlimmerM을 제공하고 있다. 사람의 경우 HGP Genome Browser와 Ensemble Genome Browser로 해석이 끝난 유전자 정보를 쉽게 검색할 수 있다.

3. 구조 예측

DNA 서열분석보다 훨씬 일찍 시작되었음에도 불구하고 단백질 구조 데이터베이스의 발전속도는 DNA 데이터베이스에 비해 훨씬 더디다. 단백질의 3차원 구조는 X-선 분광법이나 NMR 분광법으로 결정된다. GenBank에 2천만개 이상의 서열이 보관되어 있는데 반해서 PDB의 경우 2003년 10월 기준으로 약 2만 3천건 정도가 수집되어 있다.

단백질의 3차원 구조를 정확히 알면 효소의 활성자리, 반응성, 기능 등을 예측할 수 있다. 이 때문에 아미노산 서열을 이용하여 3차원구조를 예측하는 것은 분자모델링과 바이오인포매틱스 분

야에서 가장 관심을 끌고 있는 분야중 하나이다. 다만 가능한 경우의 수가 너무 많은 탓에 아주 조그만 분자를 제외하곤 현재의 기술로는 큰 어려움을 겪고 있는 분야이다.

컴퓨터를 이용하여 단백질 구조를 예측하는 모델링 방법에는 두 가지가 있다. 첫 번째는 지식기반 모델링(knowldege-based modeling)으로 이미 구조를 알고 있는 단백질의 구조 정보를 이용하는 것이다. 지식기반 모델링에는 상동성 모델링과 스레딩(threading)이 포함된다. 두 번째 방법은 주어진 정보 없이 단백질의 3차원구조를 예측하는 순이론적(ab initio) 방법이다. 순이론적 방법에서는 양자화학계산을 통해 얻어진 매개변수를 이용하여 에너지 최적화 또는 분자동력학 방법으로 구조를 예측하게 된다. 단백질 3차원 구조예측의 중요성을 보여주는 사례가 2년마다 개최되는 국제적인 경쟁인 CASP이다(http://predictioncenter.llnl.gov/). 가장 최근에 개최된 것은 2002년 11월로, 경쟁분야는 상동성검색, 스레딩, 순이론적 방법의 3개 분야였다.

RNA는 DNA처럼 유전정보를 전달할 수도 있고, 단백질처럼 효소로도 작용할 수 있는 특이한 물질이다 이 때문에 진화의 초기 단계에서는 RNA 먼저 출현하였을 것으로 짐작된다. 최근들어, 단백질로 번역되지 않고 RNA 상태에서 다양한 기능을 수행하는 non-coding RNA의 3차원 구조와 기능에 대한 관심이 커지고 있다. RNA의 구조 예측은 단백질 구조 예측보다 연구가 덜 되어 있는 상태이다.

4. DNA Microarray를 이용한 발현 분석

DNA Microarray는 주변환경 변화에 따른 전체 유전체의 발현양상을 신속하게 연구할 수 있는 방법이다. 유리판에 서열을 알고 있는 합성 DNA 조각들을 수만가지 이상 부착시킨 것으로 흔히 DNA Chip이라고도 불린다. 주변 환경변화에 따라 세포내에서 일어나는 유전자 발현 현상을 동시에 측정할 수 있고, 클러스터링을 통해 미지의 유전자에 대한 기능을 부여하는 데도 활용된다. 진핵생물에서 엑손 여부를 확인하는 데도 이용되고 있으며, 서열결정이나 다형성의 확인에도 활용될 수 있다. 마이크로어레이 데이터의 설계와 데이터 분석에 바이오인포매틱스가 크게 기여하고 있다. 이번 특집호의 다른 리뷰에서 자세한 내용을 다루고 있다.

5. 단백질체학(proteomics)

단백질체학은 특정 시간에 세포내에 존재하는 단백질 전체를 대상으로 하는 학문으로 바이오인포매틱스가 응용되는 주요 분야 중 하나로, 각 세포의 기능에 따라 생성된 단백질체(proteome)를 비교하여 질병의 원인과 주변환경변화에 따른 단백질의 정량적 분포 또는 정성적인 변화 양상을 분석하는 분야이다.

고유 전하와 질량에 따라 단백질을 분리하는 2D 젤 전기영동법, IPG를 이용한 IPG-DALT 2D, MALDI-TOF(Matrix-assisted desorption ionization-Time Of Fligh) 질량분석기 등의 실험방법을 사용하여 단백질체를 연구할 수 있게 되었다.

단백질체의 분석을 위해서는 SWISS-2DPAGE와 같은 공공 데이터베이스나 각 종에 특화된 데이터베이스를 이용한다. 펩타이드 매핑(petide mapping)에 사용되는 소프트웨어에는 PepSea, PeptIdent/MultiIdent, MS-Fit, MOWSE, ProFound, Mascot, PeptIndent2 등이 있으며, 질량분석 스펙트럼의 해석에는 SEQUEST, PepFrag, MS-Tag 등이 사용된다. 단백질체 분석을 쉽게하기 위해서 대표적인 단백질 데이타베이스인 Swiss-Prot, TrEMBL, PIR을 하나로 묶어 연합 단백질 데이터베이스 UniProt을 구축하려는 시도가 진행중이다.

VI. 바이오인포매틱스에 활용되는 인터넷 컴퓨팅과 그리드 기술

생명공학분야의 연구, 특히 단백질접힘(protein folding)과 신약설계에 필요로 하는 컴퓨팅

자원의 규모가 기하 급수적으로 증가하고 있다. 이를 해결하기 위해서 나온 개념이 바로 인터넷 컴퓨팅과 그리드 컴퓨팅이다.

1. 인터넷 컴퓨팅

인터넷 컴퓨팅이란 인터넷에 연결된 개인용 PC의 유휴시간을 이용하여 거대 계산을 분산 수행하는 것이다. 외계 생명체를 찾는 SETI@HOME로 유명해진 인터넷 컴퓨팅은 AIDS, 암, 인플루엔자, 단백질 접힘(protein folding) 등의 BT 연구분야는 쉽게 응용할 수 있다.

시스템을 기증하고자 하는 사람들은 각 인터넷 사이트에 접속하여 작은 프로그램을 다운받아 자신의 컴퓨터에 설치하면 된다. 컴퓨터의 CPU가 놀고 있을 때, 프로그램은 자동으로 해당 인터넷 사이트에 접속하여 데이터를 다운로드 받아 필요한 계산을 수행한다. 계산 종료 후에는 결과를 중앙의 서버로 모아져 전문가 그룹에 의해 해석된다. 대표적인 프로젝트로는 아래 〈표 4〉와 같다.

이외에도 ProcessTree, Great Internet Mersenne Prime Search (http://www.mersenne.org), Distribute.net 등이 인터넷 컴퓨팅 관련 프로젝트를 진행중이다. 국내의 경우 코리아앳홈(www.koreaathome.org)에서 신약후보물질의 가상탐색에 관한 사업이 진행중이다.

인터넷 컴퓨팅은 적은 비용으로 슈퍼컴퓨터급의 자원을 제공할 수 있다는 장점에도 불구하고 동시에 많은 약점을 가지고 있다. 이용 가능한 시스템의 자원량이 가변적이고, 개인용 PC를 이용하는 탓에 네트워크의 안정성 및 대역폭 등에서 취약하다. 각각의 PC간에는 직접적인 데이터 교환이 불가능하단 점 역시 약점으로 지적되고 있다. 이 때문에 데이터를 분산하여 처리하는 분야로 이용이 제한될 수밖에 없다. 이러한 단점을 극복할 수 있는 방법으로 그리드(Grid) 기술이 주목받고 있다.

2. 그리드 컴퓨팅

생물학 관련 정보는 전세계에 분포된 여러 데이터베이스에 분산되어 있음이 특징이다. 계산 자원과 데이터의 저장, 사용자 편의성 증대라는 측면에서 볼 때, 생물학에 전용된 개개의 플랫폼을 하나로 통합시킬 필요성이 더욱 커지고 있다. 그리고 분산된 DB들의 일관된 포맷의 유지, 데이터의 접근성과 보안성의 보장, 효율적인 해석을 위해서는 공통의 협업환경이 필요하다. 게다가 포스트-지놈 연구로부터 쏟아져 나올 엄청난 데이터 양은 기존의 단일 시스템으로 감당할 수 없을 정도로 커서 계산환경(데이터베이스, 프로그램 관리)과 자원에 점점 더 큰 부담을 주고 있다. 그 해결책으로 제시되고 있는 것이 그리드(grid)이다. 그리드는 데이터그리드(data grid), 계산 그리드(computational grid), 엑세스 그리드(access grid)로 분류되며 협업환경을 제공한다.

프로젝트의 목적은 전세계적으로 분포되어있는 수백 테라바이트에서 수 페타바이트에 이르는 대량의 생물학 데이터베이스의 분석과 초대형의 계산을 필요로 하는 차세대 과학연구를 가능토록 하는 것이다.

바이오인포매틱스 분야에 연구에 그리드 기술을 활용하려는 시도는 특히 유럽에서 많이 진행중이다. Eurogrid의 WP10에서는 Globus 상에서 바이오인포매틱스 응용 프로그램을 테스트

〈표 4〉 대표적인 인터넷 컴퓨팅 프로젝트

프로젝트 명	인터넷 사이트	주관 기관	목적
FightAIDS@home	www.fightaidsathome.org	The Scripps Research Institute	AIDS 치료제 개발
Folding@home	foldingathome.stanford.edu	스탠포드 대학교	Protein Folding
Parabon's Compute-against-Cancer	www.computeagainstcancer.org	NCI, Univ. W. Veriginia, U. Maryland	암 연구

하고 미들웨어의 개발 등을 진행하고 있으며, INRIA에서는 게놈관련 데이터 그리드를 구축하고 있다. EBI 역시 전세계적인 분산 데이터베이스 서비스 경험을 살려 활발하게 연구하고 있다.

1차 목표는 생물학자들에게 분산된 데이터베이스에 쉽게 접근할 수 있고 새로운 알고리즘을 테스트할 수 있는 통합환경으로 그리드를 제공하는 것이다. 이상적인 바이오 그리드를 구축하기 위해서는 데이터 형식과 시스템 관리에 있어서 표준화된 툴이 필요하다. 때문에 생물학 관련 데이터를 XML에 기초하여 통합시키려는 움직임이 활발하며,[1] 컴퓨팅 시스템과 자원의 관리에 관한 국제 표준화를 위한 워크숍이 계속해서 열리고 있다. 오브젝트 매니저, 서비스 매니저, batch 매니저, 고객 매니저 등에 필요한 기술적인 사항의 조율 역시 시급히 해결되어야 할 문제들이다.

바이오 그리드를 이룰 각각의 플랫폼들은 다음의 특징을 가지고 있다.

- 분산된 데이터 : 바이오 그리드의 기본적인 개념은 연구자간의 데이터의 실시간 공유이다. 데이터는 객체(object)형태로 그리드에 연동된 임의의 시스템에 저장된다. 물리적으로 객체가 어디에 있는지에 상관없이 사용자의 요구에 따라 객체의 추출(데이터의 추출)이 이루어진다. 데이터 사용실적을 분석하여 그리드 상에서 데이터의 재배치가 이루어지거나 미러링을 통해 추후의 시스템 효율을 높일 수도 있다. 혹시 있을지 모르는 플랫폼상의 장애를 대비하고 데이터의 신뢰도를 높이기 위해서 데이터는 몇 카피씩 복제될 수도 있다. 데이터의 미러링 시에는 실시간의 업데이트가 보장되어야 한다.
- 계산 자원 : 유전체 해석 계산을 최적화하기 위해서는 데이터 그리드 상에 연동되어 있는 CPU들을 최대한 활용해야 한다. 따라서 병렬(또는 분산)처리에 최적화되어 있는 알고리즘을 활용하여 프로세서 풀(pool)을 최대한 활용해야 한다.
- 광대역의 네트워크 : 계산 모듈간에 주고받는 데이터의 양이 엄청나기 때문에 플랫폼은 광대역폭을 가지는 네트워크에 연결되어 있어야 한다. 효율적인 네트웍 이용을 위해서는 QoS가 반드시 지원되어야 한다.
- Plug-and-Play가 가능한 해석 툴 : 필요성에 따라 계속해서 해석 툴이 추가될 것이므로 이들이 Plug-and-Play 방식으로 쉽게 바이오 그리드 시스템에 추가될 수 있어야 한다.
- 데이터의 보안성 : 그리드에 연동된 시스템 간, 혹은 외부와의 데이터 교환시 보완성이 보장되어야 한다. 특히, 각 유저별로 데이터의 접근에 차별성이 부여되어야 하고, 데이터나 시스템 제공자들이 자신의 데이터와 시스템의 보안정도를 쉽고 자유롭게 통제가 가능해야 한다.

바이오 그리드 프로젝트가 성공하기 위해서는 광대역의 네트웍 인프라와 그리드 미들웨어의 구축이 기본이 되어야한다. 그리고 실제 데이터를 생성하는 기관들의 적극적인 데이터 제공이 따라야 하고, 사용자 편의를 위한 웹 인터페이스와 좀 더 나은 알고리즘의 구현 등도 필수적이다. 여러 기간이 협력해서 진행해야 하는 그리드 사업의 특성상 기술적인 문제 못지 않게 중요한 것이 정책적인 측면이다. 국내의 경우 그리드포럼코리아(www.gridforumkorea.org)에서 관련 연구를 진행중이다.

VII. Concluding Remarks

2003년은 유전학 분야에서 아주 뜻깊은 해로 기억될 것이다. 2003년은 캐번디쉬 연구소의 왓슨과 클릭이 DNA의 이중나선구조를 처음으로 발견한지 50주년이 되는 해이다. 4월 14일에 국제 컨소시움에서 HGP의 성공적인 완료를 선언하였고, 이를 기념하여 National Human Genome Research Institute에서는 2003년 4월 25일을 'DNA의 날'로 선포하였다. 시사주간지

1) www.bioxml.org, www.bioperl.org, www.biopython.org 참조

Time에서는 DNA 관련 특별호를 발행하였고, Science와 Nature는 각각 4월 11일과 4월 24일에 특집기사를 게재하였다. Time이 1982년을 '컴퓨터의 해'로 지정한 것에 비하면 30년이나 늦었지만, 다른 한편으로 본격적인 생명공학, 아니 바이오인포매틱스의 시대가 도래했음을 인정한 것이라 할 수 있다.

30억개로 이루어진 인간의 DNA 서열을 결정하는 HGP는 지난 1990년 미국의 주도하에 6개국의 20개 센터가 참여한 국제 컨소시움에서 추진하여 왔다. 당초의 예상했던 2005년보다 훨씬 빠른 시간에 종료를 선언하기에 이르렀다. 이름과는 달리 인간의 DNA 뿐만 아니라 수많은 지구상의 생물과 질병에 대한 유전자 분석이 수행되어 온 HGP는 이는 원자를 더 잘게 쪼개고, 달에 처음으로 간 것과 비견될 만큼 역사상 가장 야심찬 과학적 업적으로 기록되고 있다.

바이오인포매틱스가 인기를 끌면서 많은 사람들, 특히 전산학 전공자들이 바이오인포매틱스 연구에 관심을 갖고 있다. 데이터베이스, 인공지능, 알고리즘 등은 전산학 분야에서 오랜 기간 다뤄온 분야이기에 쉽게 바이오인포매틱스 연구를 시작할 수 있는 것도 사실이다. 하지만 주도적인 위치를 지키기 위해서는 분자생물학과 유전학에 대한 기본적인 이해는 필수적이다. 그리고 유명 학술지에 소개되는 최신의 실험결과에도 관심을 가질 필요가 있다.

참 고 문 헌

〔1〕 "바이오인포매틱스", 한국과학기술정보연구원 심층정보분석보고서 (2003)

〔2〕 "바이오인포매틱스 관련 바이오 신기술", KOSEN 첨단 기술보고서 (2003)

〔3〕 "바이오인포매틱스", 신시아 기버스, 퍼 잼벡저, 한빛미디어 (2002).

〔4〕 김태환, 정병진, 손현석, 조영화, "바이오인포매틱스 ESTs 서열분석", 생능출판사 (2003).

〔5〕 Arthur M. Lesk, "Introduction to Bioinformatics", Oxford (2002).

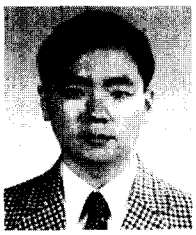
〔6〕 M. Kanehisa & P. Bork, "Bioinformatics in the post-sequence era", Nature, 33, 305 (2003).

〔7〕 A. M. Campbell & L. J. Heyer, "Discovering Genomics, Proteomics & Bioinformatics", Benjamin Cummings (2003).

〔8〕 D. E. Krane & M. L. Raymer, "Fundamental Concepts of Bioinformatics", Benjamin Cummings (2003).

〔9〕 이식, "인터넷 컴퓨팅의 생물학 응용 및 BioGrid의 동향", 슈퍼컴퓨팅소식지,(2001. 6).

저 자 소 개



이 식

1989년 2월 서울대학교 화학과 (이학사), 1993년 2월 포항공과대학교 화학과 (이학석사), 1996년 8월 포항공과대학교 화학과 (이학박사), 1994년 8월~1995년 7월 : MIT 물리학과 (방문연구원), 1996년 8월~1997년 8월 : 포항공과대학교 화학과 (연구원), 1997년 9월~1999년 3월 : Univ. of Cambridge 캐번디쉬 연구소 (연구원), 1999년 4월~2000년 10월 : Univ. of Pennsylvania 분자모델링센터 (연구원), 2000년 11월~현재 : KISTI 슈퍼컴퓨팅센터/바이오인포매틱스센터 선임연구원, 〈주관심 분야 : 바이오인포매틱스, 분산컴퓨팅, 분자모델링〉