생명과학과 컴퓨터과학

서울대학교 장병탁*
한국전자통신연구원 박선희

생명정보학(bioinformatics)은 생명과학 분야에서 생성되는 데이터를 저장, 관리, 분석, 공유하기 위한 전반적인 컴퓨터 및 정보 기술을 연구하는 학문 분야이다. 좀 더 좁은 의미로는 바이오정보(bioinformation)를 분석하기 위한 컴퓨터기술을 의미하기도 한다. 바이오정보를 다루기 위해서 컴퓨터를 활용하고자 하는 시도는 선진국에서는 이미 유전체 연구가 시작된 1990년대 초부터 본격화되었다. 그러나 국내에서 생명정보학이 많은 사람들에게 알려지기 시작한 것은 인간유전체프로젝트(human genome project)의 연구결과가 발표되기 시작한 2001년을 전후해서이다.

본 글에서는 컴퓨터과학자들이 이번 바이오정보기술 특집호를 이해하는데 도움이 될 수 있도록 분자생물학의 기본 개념과 용어를 포괄적으로 살펴보고자 한다. 그리고 현재 컴퓨터 및 정보기술이 생명과학 연구를 위해서 적용되는 분야를 간략히 알아봄으로써 뒤에 나오는 원고들에 대한 전체적인 시각을 갖는데 도움이 되고자 한다.

1. Omics의 시대

인간의 몸(body)은 1014개의 세포(cell)로 구성되어 있다. 각각의 세포는 모두 30억 개(3 x 109)의 염기(base)를 가진 DNA(deoxyribonucleic acid)로 구성된 유전체(genome)를 가지고 있다. 이 유전체는 다시 약 3만개(3 x 104) 정도의 유전자들(genes)로 구성되어 있다. 이 유전자들은 단백질(protein)을 합성하기 위한 코드에 해당하며 단백질은 우리의 몸을 구성하고 생명현상을 유지하는 기본 물질이기 때문에 결국 DNA는 생명활동을 조절하기 위한 기본 정보를 담고 있다고 볼 수 있다. DNA는 세포내에서 핵(nucleus) 안에 염색체(chromosome)의 형태로 존재하며 단백질 합성은 핵 밖에 있는 라이보솜(ribosome)에서 이루어진다[1]. 이 때 핵 내에 있는 DNA 코드가 핵 밖에서 단백질 합성에 사용되기 위해서는 이것이 전사(transcribe)되어야 하며 이 전달자(messenger)의 역할을 하는 것이 mRNA (messenger ribonucleic acid)이다. 즉 DNA의 염기서열은 mRNA에 의해 전사(transcription)되고 이것이 다시 아미노산 서열로 번역(translation)되어 단백질을 합성한다(그림 1). 이와 같이 DNA로부터 RNA가 만들어지고 이것이 다시 단백질을 생성하는 일련의 과정이 분자 수준에서의 생명현상의 기본을 이루며 이를 분자생물학의 Central Dogma라 한다.

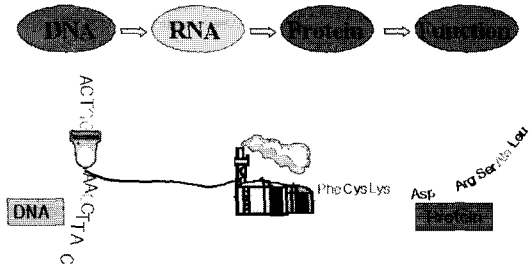

그림 1 분자생물학의 Central Dogma. DNA에 있는 유전정보는 RNA를 통해 전사(transcription)되고 이것은 다시 단백질 서열로 번역(translation)되어 생명현상을 조절한다.

위의 과정에서 DNA는 A(adenine), T(thymine), G(Guanine), C(Cytosine)의 네 개의 염기를 가진 뉴클레오티드(nucleotide)로 구성되고 RNA는 DNA와 같이 네 개의 염기로 구성되어 있으나 T 대신에 U(uracil)를 사용하여 A, U, G, C로 구성되어 있다. 단백질은 20가지의 아미노산(amino acid) 서열로 구성되어 있다. 따라서 하나의 아미노산을 지정하기 위해서는 2개의 염기로는 부족하며($4 \times 4 = 16 < 20$) 적어도 3개의 염기를 필요로 하며($4 \times 4 \times 4 = 64$) 이를 코돈(codon)이라 한다. 인간유전체프로젝트는 인간의 전체 유전체의 30억 염기서열을 분석하고 이를 바탕으로 유전자들을 찾아내고 그 기능을 밝히는 것을 목적으로 하였다. 2001년 인간유전체의 초기버전이 밝혀진 이후 유전체 후속(post-genome) 연구로서 소위 -omics의 시

* 종신회원, E-mail : btzhang@bi.snu.ac.kr

대가 도래 하였으며 그 예를 정리해 보면 다음과 같다.

- 유전체학(genomics) : 개체에 존재하는 전체 유전자들의 집합(genome)을 대상으로 하여 유전자들의 구조 및 기능을 분석하는 연구.
- 단백질체학(proteomics) : 특정 세포나 조직 또는 환경하에서 발현되는 모든 단백질들의 집합(proteome)에 대한 구조 및 기능을 대규모로 분석하는 연구.
- 전사체학(transcriptomics) : 주어진 유기체에서 전사되는 모든 전사체들의 집합(transcriptome)에 관한 연구. 유전체와 단백질체를 연결하는 역할.
- 대사체학(metabolomics) : 세포내에 있는 작은 분자 등 대사 작용(metabolism)에 관련되는 모든 물질(metabolome)에 관한 연구.
- 생리체학(physiomics) : 세포시스템이나 생화학적 시스템 또는 내분비계 등과 같은 생체 구성요소들의 생리작용에 관련하는 모든 물질(physiome)에 관한 통합적인 모델링.
- 상호작용체학(interactomics) : 세포내에서 단백질, DNA, RNA, 리간드(ligand)들의 상호작용체(interactome)에 관한 연구. 유전자망, DNA-단백질, 단백질-단백질 상호작용에 관한 연구.
- 시스템체학(systeomics) : 대사경로, 세포소기관, 세포, 생리시스템 등 바이오시스템의 구성 요소들 전체(systeome)의 상호 관계에 관한 연구.

기타 구조체학(structomics), 문헌체학(textomics), 세포체학(cellomics) 등의 용어도 사용된다. 이들 기초연구 결과들은 주로 다음과 같은 응용에 활용된다.

- 질병 진단 : 유전자 판별에 의한 질병 조기 진단
- 질병 치료 : 유전자 치료법의 개발
- 신약 개발 : 유전체학 기반 새로운 약물 유도체의 발굴
- 맞춤 의약 : 개인화된 약물 처방
- 약물 독성 분석 : 유전자 발현 패턴에 따라 약물의 독성(toxicity)을 분류
- 농생물학 : 혹한이나 가뭄 등에 강한 농작물의 육종
- 환경 : 유전체학이나 단백질체학의 연구 결과를 환경 보존에 활용
- 신소재 : 새로운 바이오 물질 기반 나노 소재의 개발

2. 생명정보학 개요

생명정보학 또는 바이오정보학은 생명 현상에 관여하는 다양한 정보를 컴퓨터를 사용하여 저장·관리·분석·해석·활용하기 위한 BT와 IT의 융합 과학 및 공학 분야로 정의할 수 있다. 이는 다시 IT 관점에서 다음과 같이 세 가지 영역으로 나누어 볼 수 있다(그림 2).

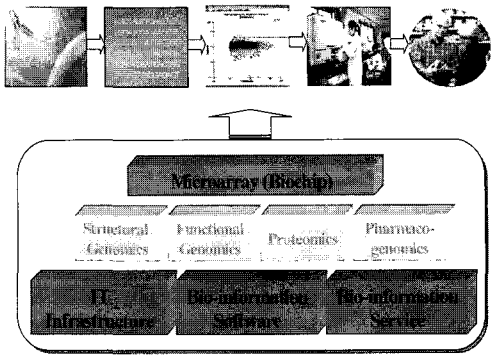

그림 2 생명정보학 관련 IT 기반 기술 및 응용 분야. IT 인프라 구축, 생물정보 소프트웨어, 바이오 정보 서비스 등의 기반 IT 기술은 구조 유전체학, 기능 유전체학, 단백질체학, 약물유전체학 등의 넓은 분야에 활용될 수 있다.

- 생명정보의 인프라 구축 : 생물정보 관리 시스템, 생명과학 연구를 위한 통신망 구축.
- 대규모 생명과학 데이터베이스의 구축 : DNA 서열, 단백질 서열, 단백질 3차 구조, 유전자 지도.
- 분석 소프트웨어 도구 개발 : 유전자 발굴, 단백질 구조 기능 예측, 유전자 조절망 구성.

컴퓨터공학의 각 분야에서 생명정보학 분야의 문제해결에 응용된 사례들을 개조식으로 살펴보면 다음과 같다.

- 데이터베이스 시스템 : GenBank, PDB 등 대규모 바이오데이터의 저장, 관리
- 서열분석 알고리즘 : DNA, RNA, 단백질 서열 데이터의 분석
- 데이터마이닝 : 바이오데이터베이스로부터 유용한 지식의 발굴
- 패턴인식 : 유전자, 인트론(intron), 모티프(motif), 마이크로 RNA (microRNA) 등의 패턴 분석
- 기계학습 : 바이오서열분석, 바이오데이터마이닝, 바이오패턴인식을 위한 기계학습 알고리즘
- 분석 소프트웨어 : 유전정보 분석, 구조 및 기능 예측 프로그램
- 프로그래밍 언어 : BioPerl, BioJava 등 새로운 프로그래밍 언어 설계
- 텍스트 정보처리 : 바이오 문헌으로부터의 정보 분석
- 온톨로지(ontology) 구축 : 바이오시스템 온톨로지

구축

- 바이오데이터 가시화 : 단백질 구조, 분자들의 상호작용망 등 데이터의 가시화
- 네트워크 인프라 : 바이오정보 서비스를 위한 네트워크의 구축
- 웹 서비스 : 정보의 웹서비스를 위한 시스템 구축
- 검색 클라이언트 : 클라이언트 서버 환경에서 바이오데이터 제공을 위한 시스템
- 컴퓨터 보안 : 바이오데이터 및 의료데이터의 공유를 위한 보안
- 수퍼컴퓨팅 : 단백질 구조 예측과 같은 대규모 CPU 시간을 요하는 문제에 활용
- 그리드 : 생명과학 환경에서의 그리드 시스템

위와 관련된 최근 연구 결과를 살펴보기 위해서는 다음의 생명정보학 관련 학술대회를 참고할 수 있다.

- RECOMB-2005, Research in Computational Molecular Biology, Cambridge, MA, May 14-18, 2005, URL : http://www.broad.mit.edu/recomb2005/
- ISMB-2005, Intelligent Systems for Molecular Biology: Annual Meeting of the International Society for Computational Biology, Detroit, MI, June 25-29, 2005, URL: http://www.iscb.org/ismb2005
- ECCB-2005, European Conferences on Computational Biology, Madrid, Spain, Sept. 28-Oct. 1, 2005, URL: http://bioinf.mpi-sb.mpg.de/conferences/eccb/eccb.htm
- BIOINFO-2005, International Conference on Bioinformatics, Busan, Korea, Sept. 22-24, 2005, http://www.ksbi.org/bioinfo2005
- PSB-2006, Pacific Symposium on Biocomputing, Maui, Hawaii, Jan. 3-7, 2006, URL: http://psb.stanford.edu/

3. 바이오정보 시스템 개발 사례

국내의 IT 기반 생명정보학 연구들이 많은 경우 바이오정보 분석 소프트웨어와 관련되어 이루어지고 있다. 여기서는 이에 대한 몇 가지 연구 사례를 간략히 소개함으로써 현재 컴퓨터 및 IT 기술이 어떻게 바이오정보를 분석하는데 활용될 수 있는지를 살펴보기로 한다. 보다 상세한 내용은 이 특집호의 각 논문에서 소개될 것이다.

- 유전자 예측 : 유전자 예측 문제는 생명체의 유전체에 존재하는 유전자의 위치를 정확하게 밝혀내는 것을 목적으로 한다. 이 정보는 유전자간의 상호관계, 유전자 산물인 단백질들 간의 상호작용, 비슷한 유전자들을 가지는 생물종간의 연관성을 밝히는데 중요한 역할을 한다[2].
- DNA 칩 데이터 분석 : DNA 칩 또는 마이크로어레이(microarrays)는 분석하고자 하는 유전자들을 칩 위에 고정시킨 후 샘플 DNA를 여기에 혼성화(hybridization) 반응시킴으로써 많은 수의 유전자들의 발현 패턴을 동시에 분석할 수 있는 새로운 기술이다. 발현 패턴의 분석을 통해 유전자들의 상관관계를 알 수 있고 이를 이용해 특정 생물학적 과정에서의 특이적인 유전자를 찾아내거나 진단 등에 활용할 수 있다. 이러한 데이터를 기계학습 기술을 사용하여 마이닝하는 기술을 여러 기관에서 연구하고 있다[3].
- 바이오 텍스트 마이닝 : 많은 생물학적 정보는 논문 형태로 발표된다. 따라서 문헌을 자동 검색하고 그 텍스트로부터 정보를 추출하여 이를 자동으로 분석할 수 있는 정보 분석, 정보 추출, 언어처리 기술 등 텍스트 마이닝 기술은 점점 그 역할이 중요해 지고 있다(그림 3). 최근 생물학적 개체명 인식, 생물학적 개체 간의 관계 인식 및 네트워크를 구성하는 연구가 전자통신연구원(ETRI)에서 진행되고 있다[4].

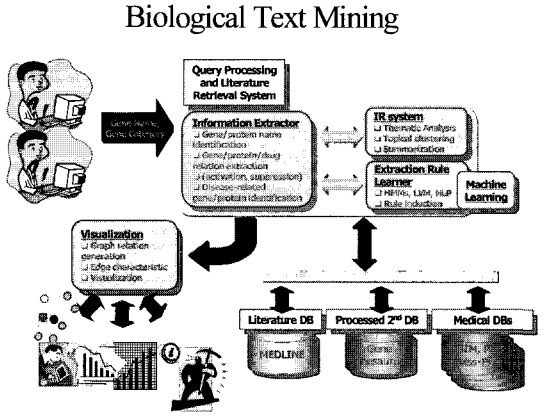

그림 3 바이오텍스트 마이닝을 통한 바이오정보 분석. 문헌 및 텍스트 데이터는 실험실 데이터 및 임상 데이터를 정제한 정보를 담고 있으며 이로부터 정보 검색, 정보 추출, 자연언어처리 기술과 기계학습 기술을 사용하여 다양하고 유용한 정보와 지식을 발굴할 수 있다.

- SNP 정보 분석 : SNP(single nucleotide polymorphism)는 서로 다른 사람들 간에 흔히 발견되는 DNA 상에서 단일 염기서열의 차이로서, 각 개인의 표현형(phenotype)의 다형성(polymorphism)이 SNP 또는 그것의 반수체(haploid)상의 조합인

반수체형(haplotype)에 따라 질병에 걸릴 확률, 성격의 차이, 다양한 재능의 차이 등을 설명할 수 있는 기반자료가 될 수 있다. 또한 SNP의 차이를 통해서 개인의 유전자 특성에 적합한 약을 처방한 다든지 치료 방법을 달리 할 수 있어 개인화된 의료 서비스(personalized medicine)가 가능해 질 수 있다. SNP를 분석하고 이를 질병 치료에 활용하기 위한 데이터베이스가 국가유전체정보센터를 통해서 이루어지고 있다[5].

- 의료정보학 : 임상 의료 데이터를 효율적으로 습득, 관리, 검색, 응용하기 위해 컴퓨터 및 정보기술을 활용하는 연구로서 이미지데이터의 분석, 지식관리 시스템, 데이터베이스 기술, 네트워크기술 및 컴퓨터하드웨어 등 많은 분야가 관련된다. 특히 최근에는 유전체 및 바이오정보를 활용하여 질병 진단과 치료 등에 활용하려는 시도가 있다[6].
- 시스템 생물학 : 신호전달(signal transduction), 세포 주기(cell cycle), 세포 분화(cell differentiation), DNA 복제(replication), 물질대사(metabolism) 등과 같은 세포의 생리활성 반응 조절에 관여하는 단백질들 또는 유전자 간의 상호작용 망을 구성하고, 그 특성을 연구하며, 동역학적(dynamics)인 관점에서 각 분자의 양과 활성정도를 모델링 하는 연구이다. 최근 재구축된 유전자, 단백질 망을 이용하여 신약 물질의 후보를 발굴하는 연구도 각광을 받고 있다. 이러한 연구에 대한 소개가 [7]에 포함되어 있다.

4. 교육, 연구, 사업적 측면

인간유전체 프로젝트와 단백질체 프로젝트 등 생명과학의 발전은 컴퓨터공학과 IT 기술에 새로운 도전과 기회를 제공해 준다. 전에 접하지 못했던 새로운 데이터와 문제를 새로운 형태로 제시하는 점에서 새로운 기술 개발의 수요를 창출하고 있으며 이는 기존의 컴퓨터 회사들에 새로운 사업 영역으로 등장하였다. IBM, Microsoft, HP, Oracle, Siemens 등 세계적인 컴퓨터 회사들 뿐만 아니라 국내 기업으로서는 Samsung 등도 생명과학(Life Science)을 새로운 사업 분야로 명확히 분류하고 있다[8]-[14]. 또한 선진국의 여러 대학에서 이러한 시장 및 산업적 수요를 충족시키기 위한 학제적인 교육 및 연구 프로그램을 신설하고 있다. 그 예로서 MIT의 CSBI 프로그램[15], Stanford 대학의 Bio-X[16], Princeton 대학의 Integrative Genomics 프로그램[17], Harvard 대학의 Bioinformatics and Integrative Genomics 프로그램 [18] 등이 있다.

국내에서도 많은 대학들이 지난 5년 동안 학부나 대학원 과정의 프로그램들을 신설하였다. 부산대와 숭실대에서 처음으로 각각 대학원과 학부 과정의 프로그램을 개설하였으며 현재 서울대, 고려대 등에 생명정보학 관련 대학원 협동과정들이 개설되어 있으며 KAIST에서는 바이오시스템 학과가 문을 열었다. 산업적인 측면에서 국내 생명정보학 전문 회사의 수는 아직 그리 많지 않다[2, 19]. 그리고 이들이 아직 명확한 사업 모델을 갖고 있지 못한 것도 사실이다. 기술의 특성상 IT 기술만으로 보다는 BT와 접목되어야 하는데 이를 뒷받침할 학제적인 교육을 받은 고급 연구자의 수가 아직 절대적으로 부족한 형편이다.

생명정보학을 연구하는 데 있어서 중요한 것 중의 하나는 이것이 학제적인 연구 기술 분야라는 것을 인식하고 서로 협력 연구하려는 의지와 신뢰를 기반으로 하는 것이다[19]. 생명과학자들에게 생명정보학은 도구 학문의 성격이 강하다. 다른 응용 소프트웨어의 사용자들이 그러하듯이 생명과학자들은 생명정보학으로부터 많은 데이터를 빨리 효과적으로 분석하고 가시화할 수 있으며 편리하게 사용할 수 있는 소프트웨어를 기대한다. 분석의 기반이 되는 알고리즘과 기술적인 주제는 그 결과가 흥미롭다는 조건하에서나 관심이 있다. 반면, 컴퓨터공학자들에게는 바이오데이터 분석을 통해서 새로운 IT 기술을 개발하고자 하는 욕심이 있다. 예를 들어, 바이오데이터는 차원이 높으면서 샘플의 수는 적고 노이즈가 심한 데이터를 분석해야 하는 새로운 문제를 던져준다. 또한 데이터베이스 관점에서 웹 상에 널리 흩어져 있는 초대규모의 혼성 데이터들을 통합 분석해야 하는 새로운 문제를 제시한다. 그러나 이러한 기술적으로 흥미있는 문제는 종종 생명과학자들에게는 흥미 있는 문제가 아닐 수 있다. 이러한 두 분야의 관심 및 요구 조건들이 합의를 이루었을 때에 협력 연구가 잘 이루어지고 시너지 효과가 발행하며 경제산업적으로도 효과가 큰 부가가치 높은 새로운 기술이 탄생하게 될 것이다.

마지막으로 **본 고에서는 좁은 의미에서의 바이오정보기술인 생명정보학에 대하여 다루었다. 이는 BT 연구개발을 위해 IT 기술을 활용하는 것인데 반해서 BT 기술을 활용해서 새로운 IT 기술을 개발하기 위한 바이오정보기술에 관한 연구도 있으며 이에 관한 사례로서 이번 특집호의 [20]을 참고하기 바란다.** 보다 포괄적인 생명과학에 기반한 컴퓨터과학에 관한 BIT 연구에 관해서는 [21]과 [22]를 참고 할 수 있다.

참고 문헌

[1] Brown, T.A., Genomes, Bios Scientific

Publishers, 2nd Ed., 2002.
〔2〕 박기정, "유전체학과 생물정보사업", 제23권 제5호, pp.24-31, 2005.
〔3〕 박성배, "기계학습/텍스트마이닝과 생명과학", 제23권 제5호, pp.32-40, 2005.
〔4〕 박선희, "정보통신부 BIT관련 연구개발 사업 현황/계획", 제23권 제5호, pp.57-69, 2005.
〔5〕 박종화, "국가유전체정보센터(NGIC): 생명정보사회를 대비한 10만 양병", 제23권 제5호, pp.76-81, 2005.
〔6〕 김철민, "유전체의학을 위한 생명의료정보학", 제23권 제5호, pp.10-17, 2005.
〔7〕 조광현, "시스템생물학 연구동향", 제23권 제5호, pp.18-23, 2005.
〔8〕 IBM, IBM Healthcare and Life Sciences, http://www-1.ibm.com/industries/healthcare/
〔9〕 Microsoft, Microsoft Healthcare and Life Sciences, http://ww.microsoft.com/industry/healthcare/
〔10〕 Hewlett-Packard, HP Life Sciences Initiative, http://www.hp.com/techservers/life_sciences/
〔11〕 Oracle, Life Sciences Platform, http://www.oracle.com/technology/industries/life_sciences/
〔12〕 Siemens, Medical Solutions, http://www.medical.siemens.com/
〔13〕 Samsung, Healthcare Products, http://www.samsungamerica.com/Text/healthcare.asp
〔14〕 Samsung, Samsung Advanced Institute of Technology (SAIT) Digital Bio Lab, http://www.sait.samsung.co.kr/sait/saitBioLabIntro.jsp
〔15〕 MIT, Computational and Systems Biology Initiative, http://csbi.mit.edu/
〔16〕 Stanford University, Bio-X, http://biox.stanford.edu/
〔17〕 Princeton University, Integrative Genomics, http://www.genomics.princeton.edu/
〔18〕 Harvard University, Bioinformatics and Integrative Genomics, http://big.chip.org/
〔19〕 윤정호, "국내 바이오산업과 BIT 연구개발 현황", 제23권 제5호, pp.70-75, 2005.
〔20〕 장병탁, "나노바이오지능 분자컴퓨터: 컴퓨터공학과 바이오공학, 나노기술, 인지뇌과학의 만남", 정보과학회지, 제23권 제5호, pp.41-56, 2005.
〔21〕 장병탁, "컴퓨터 바이오과학과 바이오 컴퓨터과학", 정보과학회지, 제21권 제6호, pp.5-13, 2003.
〔22〕 Forbes, N., *Imitation of Life*: How Biology Is Inspiring Computing, MIT Press, 2004.

장 병 탁

1986. 2 서울대학교 컴퓨터공학(학사)
1988. 2 서울대학교 컴퓨터공학(석사)
1992. 7 독일 Bonn 대학교 컴퓨터과학 박사
1992. 8~1995. 8 독일국립정보기술연구소(GMD) 연구원
1995. 9~1997. 2 건국대학교 컴퓨터공학과 조교수
1997. 3~현재 서울대학교 컴퓨터공학부 부교수, 인지과학, 뇌과학, 생물정보학 협동과정 겸임.
2003. 8~2004. 8 MIT Computer Science and Artificial Intelligence Laboratory(CSAIL) 객원교수
2001. 1~현재 바이오정보기술연구센터(CBIT) 센터장
2002. 6~현재 바이오지능 국가지정연구실 실장
관심분야 : Biointelligence, Probabilistic Models of Learning and Evolution, Molecular/DNA Computation

박 선 희

1976~1981 서울대학교 수학교육(학사)
1982~1986 Univ. of Texas(Austin) 수학(석사)
1986~1989 Univ. of Texas(Austin) 물리학(박사)
1990 Center for Relativity at Univ. of Texas, Postdoc
1990~1991 I.C.T.P.(Italy) Postdoc
1991~1994 Center for Theoretical Physics at S.N.U. Postdoc
1994. 8~현재 한국전자통신연구원 연구원
관심분야 : 바이오인포매틱스, 생체정보처리