식품산업과 영양 5(3), 47~51, 2000 Food Industry and Nutrition

특집 : 한방생약자원의 식품·생명산업적 이용

생약자원 연구에 있어서의 생물정보학의 활용
Medicinal Herb Research and Bioinformatics

김 상 배 (Sangbae Kim)
(주)파라바이오

서 론

지난 수년동안 생물학, biotechnology, 의/약학, 농학, 환경공학과 같은 생명체와 관련된 여러 분야들에 있어 크나큰 발전이 있었다. 발전의 방향은 크게 두 가지 핵심분야로 대변되는데, 하나는 다양한 형태의 automated tool을 이용하여 생물체로부터 다양한 종류의 data를 도출하는 분야이며 또 다른 하나는 이렇게 도출된 massive data를 컴퓨터로 처리하여 여러 가지 유용한 지식을 얻어내어 활용하는 것이다. 수 십년전부터 태동하기 시작하여 현재까지, 기초 생명과학의 주된 연구 방식이 되어왔던 molecular biology는 각각의 개체가 가지고있는 관심 있는 요인들을 개별적으로 분리하고 이들의 성질과 control mechanism을 밝히는 방식이었다면, 최근의 'genomics'적인 연구방식에서는 대상이 되는 생물체로부터 가능한 한 최대한의 정보와 이를 구성하는 요인들을 도출시켜 이들이 어떠한 complexity를 가지며 서로 어떤 상관관계 하에서 생명체를 control하는 mode of action을 가지고 있는가를 전체적인 시각에서 다루어 보고자 하는 방식이다.

흔히 생명현상이라는 것은, 단편적이고 local한 정보로 해석되는 경우에는 여러가지 confounding factor들에 의해 잘못 해석되는 경우가 많으므로 가능한 한 많은 요소들을 종합하여 이들 각각의 요소가 어떠한 networking을 이루어 특정 현상을 나타내는지를, 전체를 살펴보고 다룰 수 있을 때에만 진정으로 그 현상을 이해할 수 있을 것이다. 이러한 것이 가능할 때 비로소 우리는 이를 조정하여 우리에게 필요하고 유용한 결과를 도출시키고 나아가 잘못된 부분을 수정하는 등의 일들을 수행할 수 있게 되는 것이다. 이를 위한 노력의 시발점이며 가능성을 제공한 것은, 1990년도부터 미국을 필두로 하는 선진국들을 중심으로 본격적인 연구가 시작되어 이제 그 첫 단계 목표, 즉 인간의 유전정보 전체를 밝히는 작업을 조기 마감한, 인간 유전체 프로젝트(human genome project)라고 할 수 있다.

Human genome project는 실제로 여러 분야에 막대한 영향을 주고 있으나, 국내의 상황을 살펴보면 소수의 관련 분야에 있는 사람들을 제외하고는 이 project가 미칠 영향에 대한 대국민 인식이 매우 저조한 것이 사실이다. Human genome project가 미치는 영향은 학문적으로 이용할 수 있는 인간 유전자에 관한 데이터를 얻었다는 단순한 데 있는 것이 결코 아니라 이 project를 근간으로 신종의 산업이 형성될 것이라는 예상이다. Human genome project에서 사용된 새로운 패러다임의 연구방식, 즉 automated tool과 computer based technology를 결합한 research method는 하나의 출발점이었으며 이제는 하나의 예에 불과한 것이 되었다. 더욱 중요한 사실은 생명체와 관련된 여러 다른 수많은 기존의 연구개발 방식을 완전히 바꾸는 혁신적인 방식이 되었다는 사실이다. 이로 인해 국내외에서 수많은 R&D를 위주로 하는 바이오텍 회사들이 생겨나게 되었으며 특히 인간의 질병대응 소재나 제품을 개발하고자 하는 목적의 회사들이 많다. 이러한 신소재시장은 그 규모가 반도체나 IT 시장보다 훨씬 더 팽창하리라 예상되고 있다.

다가오는 21세기에는 이러한 바이오텍 산업이 세계경제의 주역이 될 것이라고 모두들 예상하고 있으며 이러한 세계적 동행에 우리도 선진국의 대열에 동참을 할 수 있기 위해서는 이러한 연구동행에 대해 이해하고 동참하는 의식이 절실히 요구되는 시점이다. 이 글에서는 이러한 새로운 연구의 패러다임의 하나인 bioinformatics(생물 정보학)에 관해 개괄적인 내용을 소개 함으로서 high tech.보다는 우리에게 더욱 경쟁력이 있는 한방생약 분야에 종사하는 사람들에 있어 세계적인 연구방식 동향과 향후에 이들을 활용할 수 있는 의식을 공유 하고자 한다.

Information derived from living organisms

모든 생명체의 세포핵에는 긴 나선형의 정보기록장치인 염색체DNA에 유전정보라는 blue print를 가지고 있으며 4가지 염기의 조합을 통해 정보가 기록되는 이 DNA 속에는 합계 약 60억개의 염기를 가지고 있으며, 기록장치에 coding 되어 있는 사람의 유전자의 수는 약 10만개 정도가 된다. 이 10만 가지 유전자가 transcription을 거쳐 RNA 분자가 되어 cytosol로 방출된 후, translation을 거쳐 다시 protein 이 만들어진다. 이 단백질은 다시 다양한 형태의 post-translational modification을 거친 후 folding이 되고

같은 종류의 동일 과정에 관여하는 일련의 단백질끼리 복합체를 이루게 된다. 결국 모든 생명현상은 이러한 최종 단백질 생성물과 그들의 복합체의 복잡한 상호작용의 결과라고 말할 수 있다.

이러한 유전자를 비롯한 세포라는 시스템을 구성하는 요소들은 서로 상호작용에 의해 특정반응을 나타내게 되는데 이것이 결국 하나의 세포가 나타내는 성질이 된다. 이러한 세포의 반응적 성질은 외부에서 입력되는 정보와 그 자체의 판단과 필요에 의해 다양한 반응과 변화를 보여주게 되며 동일한 종류들이 다수 결합하여 조직이 되고 이러한 조직들이 다수 모여 특정 작업을 담당하는 organ을 형성하게 된다.

우리가 생명현상을 이해하기 위해서는 해독과 조작의 작업이 필요하다. 우리가 생명체로부터 표현되는 생명현상을 하나의 시스템으로 본다면, 흔히 시스템의 분석을 위해서는 그 시스템의 구성요소 들을 먼저 파악해야 한다. 요소들을 찾아낸 다음에는, 요소들의 연결과 그 시스템에서 표현되는 성질에 대한 정보를 해독해내야 한다. 생명체의 경우 하나의 세포가 가지는 구성요소들은 수만에 달하게 되고, 이렇게 하여 얻어지는 정보는 실로 막대한 양이 된다. 신뢰성 있는 통계적 수치를 보여주기 위해서는 다시 이러한 작업 전체를 다수의 다른 조직, 기관 그리고 개체로부터 얻어진 것에 대해 반복하는 형태가 되어야 하는데 일반적으로 한 가지 실험가설에 대해 수집되어야 하는 데이터 포인트의 수는 쉽사리 수십 억을 헤아리게 된다. 이와 같은 방식을 도입한 생물체에 대한 새로운 연구방식의 하나의 예로 genomics를 들 수 있다. Genome이란 어떤 한 생물체가 가지는 유전정보의 '전체'를 뜻하는 것으로서 genomics란 어떤 특정 생물체가 가지는 유전정보와 이로부터 발현되는 전부를 총체적으로 연구하는 학문을 뜻한다. Poteomics도 위에 설명한 소위 genomics적인 것들과 본질적으로는 아무런 다를 바가 없다. 단지 RNA의 양들을 추적하는 대신에 좀 더 복잡한 도구로 더 복잡한 성질의 단백질들에 대한 데이터를 얻어내는 정도의 차이가 있을 뿐이다.

DNA microarrays

DNA microarrays는 흔히 DNA chip이라 불리우기도 하며 수 cm 또는 그 이하의 유리 또는 실리콘 등으로 된 칩 위에 수만에서 수십만 가지에 달하는 서로 다른 DNA 분자를 2차원 매트릭스 형태로 심어놓은 것으로서 DNA 분자가 pico mole 또는 femto mole 단위로 집적되어 생물체로부터 정보를 얻어내는 데 쓰이는 현재 가장 각광을 받고 있는 분석방법 이다. 이렇게 만들어진 DNA 칩은 생물체로부터 얻어진 시료와 반응을 시키게 되는데, 이때 사용되는 시료는 DNA 또는 RNA이며, 보통 형광물질로 표지가 되어 있다. 즉, 어떤 생물체 시료에 약물의 투여와 같은 조작을 가하거나, 또는 암조직과 같은 것에서 DNA나 RNA를 추출하여, flurolescence tag을 붙여 수용액 상태의 적절한 조건에서 이 칩 표면에 treatment한 시료와 hybridization을 하도록 한다. hybridization signal의 intensity가 바로 얻고자 하는 데이터가 된다. 일단 여기까지 과정에서 두 가지 종류의 중요한 전산적인 도구가 존재하는데, 하나는 signal (그리고 image) processing에 관한 것이며 다른 하나는 단 한 종류의 칩에도 수만 가지 또는 수십만 가지의 서로 다른 DNA 염기열이 붙여야 하는 것이므로 이들의 분석을 위해 전산도구가 필요하게 된다. 그리고, DNA 칩의 장점은 하루에도 수천 번이나 수만 번, 또는 그 이상을 해낼 수 있다는 데에 있다. 이에 수반되는 물리적인 리소스의 수만 해도 엄청나므로 이를 관리하고 추적할 수 있는 전산적인 도구 또한 매우 중요하다. 이는 소위 LIMS(Laboratory Information Management System)라 부르며, 이러한 연구방식의 기반을 형성하는 중요한 전산적인 도구이다. 이러한 데이터로부터 무언가를 알아내는 작업을 data mining이라 부른다. DNA 칩으로부터 얻어진 데이터에 대해서는 현재 clustering을 해보는 수준 정도에 머무르고 있으나, 앞으로 더욱 발전적인 기법들이 등장하여 더 많은 결과들이 쏟아져 나오기 시작하면 지금까지 우리가 생명현상을 전체로 보지 못하고 지엽적인 즉, 장님 코끼리 만지기 식의 연구 방식이었는지를 깨닫게 해줄 것이다.

Bioinformatics란?

생물정보학은 영어로는 bioinformatics, computational biology, 또는 computational molecular biology 등의 용어로 불리는 분야로서 국내에서는 생물정보학이라는 용어로 어느 정도 확립이 되어가고 있는 상황이다. 생물정보학은 매우 다양한 분야를 담고 있는 폭넓은 것이며, 굳이 정의를 내린다면 '분자 생물학(biology in terms of molecules)에 정보학 기술을 접목하여 이러한 유전적 또는 생물학적 정보를 이해하고 관리하는 학문' 혹은 '생명현상 연구에 필요한 다양한 전산학/통계학/수학적인 것들'이라는 표현이 그나마 본질에 어느 정도 접근을 하는 것이라 할 수 있다.

두 가지 중요한 생물정보학의 대상이 되는 데이터는 DNA와 이에 코딩된 정보로부터 만들어지는 단백질의 서열정보이다. 또 한 가지 중요한 생물정보학의 대상이 되는 데이터는 바로 structural biology에서 다루어지는 1차원적인 아미노산 서열로부터 folding에 의해 3차원적인 단백질을 예측해 내고 3차원적인 형태를 상호 비교하며 주어진

단백질에 ligand와 translation factor로 작용하는 물질들을 찾아 내는 등 매우 다양한 문제들이 있게 된다. 그 다음은 이러한 단백질들과 염색체 상에 DNA 형태로 들어있는 유전자, 그리고 이들 사이의 중간 단계라 할 수 있는 RNA가 어떤 것들끼리 어떻게 상호작용을 하며, 어디에 얼마나 존재하고, 어떤 환경이나 조건에서 어떻게 양이나 구조 등이 변하는지에 대한 데이터가 있다. 이들을 밝혀내는 현재 가장 각광을 받고 있는 도구가 바로 앞서 소개한 DNA 칩과 proteomics라 불리는 것들이다. DNA 칩 또는 proteomics 등은 외부에서 연구의 디자인에 따라 다양한 treatment 혹은 변이를 준 상황에서 가능한 한 많은 양의 데이터를 도출하여 이들의 결과를 분석함으로써 생체내에서 실제로 일어나고 있는 생명현상을 밝혀내고자 하는 것이 결국 genomics 또는 proteomics의 활용 목표이다.

이런 과정을 거쳐 획득된 데이터는 결국 복잡한 여러가지 요소들의 상호작용에 의해서 발생된 복잡한 현상의 한 단면이다. 단순한 관계해석으로는 이러한 상호작용을 이해할 수 없기 때문에 이들을 해석하기 위해서는 통계적인 처리나 복잡한 modeling 방법을 사용할 수 밖에 없다. 이 점이 바로 오늘날의 생물정보학의 태동의 이유이며 생물정보학의 가장 큰 부분을 이루고 있는 핵심적인 내용이다. 그리고, 이러한 모든 것들에는 자동화가 매우 중요한데, computational method를 이용하여 manual한 방법이 아니라 자동화된 장치의 사용이 가능해 진 것이 바로 genomics 혁명이 도래한 하나의 큰 배경이 된 것이다. 단순히 기계의 자동화된 조작만이 아니라, 각 단계에서의 일어나는 일들에 대한 해석의 자동화도 매우 중요한 점 중의 일부이다. 왜냐하면, 비록 각 단계에서 대량의 데이터를 얻어낸다 해도, 그들 사이의 연결 부분에서 사람이 일일이 해석과 판단을 해주어야 한다면 결국 down-time이 생겨 전체적으로 비효율적인 작업이 될 수 밖에 없을 것이다. 이 의사결정도 bioinformatics분야의 큰 한 축을 이루는 중요한 부분인 것이다. 또 한가지 중요한 것은 genomics와 proteomics 모두에서 가장 최초로 얻어지는 데이터는 결국 image 형태의 데이터로 나타나는 것이 대부분이다. 그러므로, image analyzing process도 매우 중요한 분야의 하나이다. 이처럼 소위 bioinformatics라 불리는 분야는, 온갖 다양한 computational/statistical/mathmetical/physical적인 것들을 모두 담고 있는 총체적 분야라 할 수 있다.

생약자원 연구에 있어서의 bioinformatics의 활용

Bioinformatics라는 분야는 실제로 매우 폭 넓은 분야이다. 단순한 생물학적 data base로부터 prognosis를 위한 신소재를 screening하는 대규모 epidemiological study까지 매우 다양하게 존재한다. 단지 IT의 support가 필요하다는 점이 특별하다. 그리하여 bioinformatics라 불리우며 단지 data handling외에 IT support에 의한 좀 더 까다롭고 차원 높은 분석도 가능한 분야라 이해하면 된다. 사용자들은 이러한 bioinformatics를 활용하여 필요한 정보를 도출해 내는 것이 더 현실적인 설명이 될 것이다. 생약자원 연구에 있어 bioinformatics를 활용하는 방법에는 여러가지의 다양한 접근 방법이 있으며 Table 1의 요약을 참고로 하기 바란다. 최근에는 web환경의 활용으로 bioinformatics활

Table 1. Bioinformatics related data bases

Clasification	Contents	
Biological database	Nucleic acid sequence	EMBL
	Protein sequence	SWISSPORT
	Protein sequence	PIR
	EST sequence database at NCTI	DbEST
	Non-redundant protein sequence	OWL
	General structural databases	PDB
	Organism specific database	C.Elegans at the Sanger Center
	More specialized data	KEGG
Simple database queries	Simple queries of many databases	SRS at the EBI
Sensitive sequence analysis	Diagnostic patterns for protein	PROSITE
	Slightly longer patterns	PRINTS
	Similar to PRINTS	BLOCKS
	Profile hidden Markov models	Pfam
	Iterative database searches	PSI-BLAST
Prediction of 3D structure	3-D structure/protein folding	SWISSMODEL
	Prediction of secondary structure	JPred
		Predict Protein
	Prediction of transmembrane segments	PredictProtein
Genome analysis Gene prediction	Comparison of protein w/genomic DNA	The WISE2 package
	Splice site	FGENEH at the Sanger Centre

용에 실로 획기적인 변화가 일어나 그 사용이 매우 쉬워져 대부분의 작업이 data base나 software를 download받지 않고도 web상에서 직접 service를 받을 수 있게 되었다. 중요한 site로서는 The European Bioinformatics Institute, The N.C.B.I., 그리고 ExPASy등이 있다.

생약자원의 screening에 있어 중요한 활용분야의 하나는 바로 functional ingredient control in gene expression의 분야라 하겠다. Screening하는 소재의 mode of action을 밝히는 과정에 있어 후보 소재가 특정 증상/질병에 관여하는 유전인자의 transcription, RNA stability/translation/ protein degradation/post translational modificatio의 일련의 과정에 미치는 영향을 database와 연계하여 bioinformatics의 활용으로 설명이 가능할 것이다(Fig. 1). 특히 생약소재에 함유된 polyphenol성분이나 bezene ring compounds들은 특히 유전인자의 발현에 관여하는 여러가지 factor들로 작용하게 되므로 이러한 연구에 활용도가 매우 높다고 생각된다.

좀 더 이상적인 활용분야를 살펴보면, DNA 칩은 한 생물체의 genome 전체 또는 일부에 대한 대규모 분석, 조직이나 세포로부터 발현되는 유전자 전체의 종류와 양을 측정할 수 있는 도구이다. 예를들어 일정 수(많으면 많을수록 좋다)의 60세 이상 고혈압 환자로부터 DNA를 추출하여 이를 위해 특별히 제작된 칩과 반응시킨 데이터를 얻어내고, 다시 60세 이상 고혈압이 없는 사람들의 DNA를 추출하여 역시 칩과 반응시킨 데이터를 얻어낸다. 물론 이렇게 도출된 시료와 이에 수반되는 다양한 정보들을 다루기 위해서는 bioinformatics의 활용이 중요한 역할을 하게 된다. 이렇게 대량의 데이터가 있을 경우에는 이 데이터를 이용하여 이들이 공통적으로 가지는 성질을 반영하는 모델을 컴퓨터 내부에 만들 수 있다. 즉, 일반적인 경우에는 아직 고혈압과는 거리가 먼 30대(또는 20대)로부터 DNA를 추출하거나 특정 생약제재를 투여한 모델과 맞추어 보

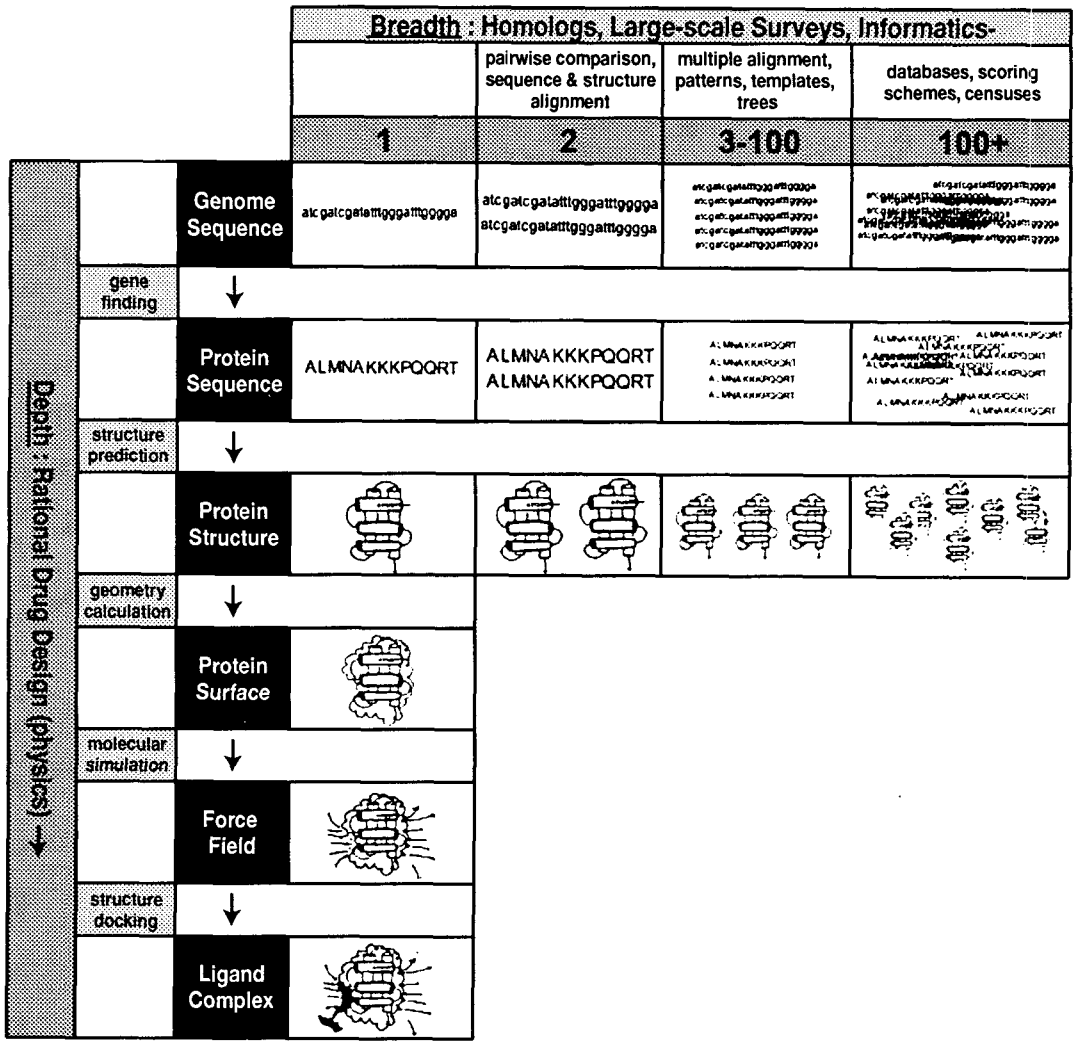

Fig. 1. The bioinformatics spectrum.

는 작업을 하면, 그 사람이 장차 고혈압으로 고통받게 될 확률이 얼마나 되는지 혹은 약재가 어떤 특정 인자에 변화를 주어 증상을 예방하는가등의 정보를 얻을 수 있을 것이다. 이와 같은 방식을 diagnosis와는 달리 prognosis라 부르며, 앞으로 큰 시장을 형성하게 될 것이라 예상되고 있다. 또한 이것이 바로 왜 DNA 칩과 bioinformatics가 붐을 이루며, 대형 제약회사들이 앞다투어 투자를 하고 있는가에 대한 극히 일면에 불과하지만 쉽게 이해가 가는 예라 할 수 있다. 그리고, 이것은 최소한 7만개는 넘을 것으로 예상되는 인체 유전자 개개의 자세한 작용기작을 모두 이해하게 될 수년 내로 가능해 보이는 응용 분야가 바로 눈앞에 다가와 있는 것이기 때문이다.

결 론

이상에서 생물정보학 분야의 최근 동향과 생약자원 연구와의 연계에 대해 간략히 살펴보았다. 유전자에 대한 연구 또한, 이미 수십 개로 그 수가 늘어났고 앞으로 그 발전 속도는 우리들의 상상을 초월할 정도로 급속히 발전해 나갈 것이며 전체 genome의 유전정보가 알려진 생물체들의 비교와 종합을 통해 새로운 양상으로 또한 발전하게 될 것이다. 어느 한 생물체에 속한 하나의 유전자, 단백질, 또는 그 유전자에 관여하는 기능성 소재등은 개별적인 연구의 대상이 아니라, 생물계라는 거대한 조직을 구성하는 일원으로서 서로 밀접하게 얽힌 동적인 변화의 과정으로 이해 되어야 할 것이다. 물론 이 모든 emerging research method에 있어서 정보기술이 중요함은 의심할 여지가 없다. 이러한 생물정보학의 발전은 단순히 그것을 실행할 수 있는 도구만 있으면 되는 것이 아니다. 우리가 예를 들어 생약자원의 연구에 있어 일정 사용료를 지급하고 database를 활용(초기에는 이런 방법으로 접근하겠지만)하는 등의 소극적인 자세에서 향후에는 생명현상에 대한 연구 그 자체를 위해서 연구에 대한 이해와 인력의 양성도 필요하다. 이러한 관점으로 볼 때 생약자원을 연구하는 분야에 있어서도 이러한 emerging technology와 동향에 대한 더욱 활발하고 적극적인 동참과 연구infra를 조성하는데 있어 속도감을 가지고 의식을 공유하는 적극적 자세가 요구되는 시점이다.

감사의 글

본 고를 작성 하는데 많은 도움을 주신 생물정보연구소의 원세연 박사님께 심심한 감사의 말씀을 드립니다.

참 고 문 헌

1. Gershon, D., Sobral, B.W., Horton, B., Wickware, P., Gavaghan, H. and Strobl, M. : Bioinformatics in a post-genomics age. *Nature*, **389**, 417-422 (1997)
2. Smith, T.F. : Functional genomics-bioinformatics is ready for the challenge. *Trends Genet.*, **14**, 291-293 (1998)
3. Brutlag, D.L. : Genomics and computational molecular biology. *Curr. Opin. Microbiol.*, **1**, 340-345 (1998)
4. Li, Z.R., Tian, A.J. and Yang, Y.Y. : Preparing for the third millennium: the views of life informatics. *Medinfo.*, **1**, 394-396 (1998)
5. Altman, R.B. : Bioinformatics in support of molecular medicine. Proc. AMIA Symp., p.53-61 (1998)