바이오 정보 기술

Bioinformatics Technology

정보통신 미래기술 특집

정호열 (H.Y. Jung) 바이오정보연구팀 선임연구원
박수준 (S.J. Park) 바이오정보연구팀 선임연구원
박선희 (S.H. Park) 바이오정보연구팀 팀장

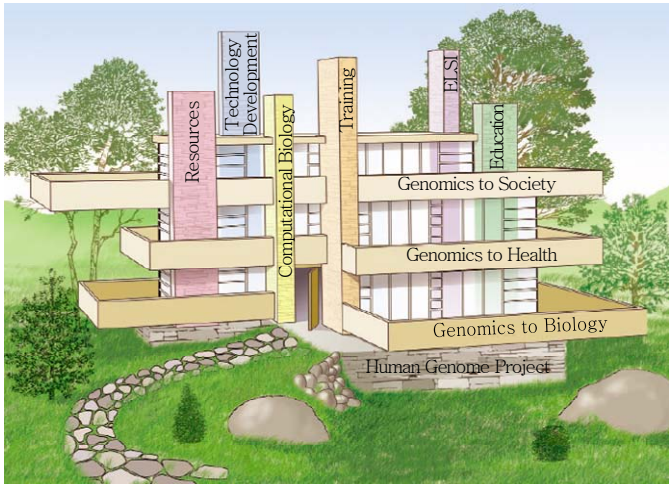
목 차

Ⅰ. 개요
Ⅱ. 바이오 정보 기술 동향
Ⅲ. 바이오 정보 시장 동향
Ⅳ. 결론

현재 우리나라에서 가장 주목받고 있는 분야가 IT와 BT일 것이다. IT는 워낙 언론매체에 많이 노출되어 어떤 것을 말하고 어떤 분야가 있다는 것을 많은 사람들이 잘 알고 있는 실정이나, BT 관련해서는 단순히 시험관을 연상하는 사람이나 줄기세포 연구 정도로 알고 있는 사람들이 많을 것이다. 현재의 BT 분야는 예전의 소규모 실험을 벗어나 대규모의 실험 수행이 가능한 시스템이 구축되었는데, 이러한 대규모 실험 결과를 분석하기 위한 정보학적인 방법의 도입이 필수적인 시대가 되었다. 그래서, 이런 접근 방법을 통상 IT와 BT의 융합기술이다라고 이야기한다. 바이오 정보 기술이란 이런 대규모의 생물학적 데이터를 시스템적으로 분석하여 정보를 추출하는 제반 기술이라고 이야기 할 수 있다. 일반적으로 많이 알려진 용어로는 생물정보학(바이오인포매틱스) 혹은 계산생물학이 있다. 더 넓은 의미에서 이야기 할 때는 이러한 정보 추출을 위한 분석과정 뿐만 아니라, 생물학 관련 데이터베이스 구축 및 서비스 부분도 포함해서 이야기하고 있다.

Ⅰ. 개요

바이오 정보 분석은 일반적으로 생물정보학(바이오인포매틱스, bioinformatics) 혹은 계산생물학(computational biology)이라고 이야기한다. 생물학적 실험이 대규모로 이루어지기 시작하면서 과거의 소규모 분석으로는 더 이상 생물학적 데이터를 처리하는 것이 불가능하여졌다. 대표적인 것이 생물체의 DNA 서열을 판독하는 것인데, 사람의 염기서열은 30억 개(생물학에서는 base라고 함) 정도 되는 것으로 알려져 있는데 이전의 생물실험에서 몇 천 베이스 정도는 실험자가 눈으로 보면서 하는 것이 가능하였지만, 현재의 30억 베이스 정도의 규모를 실험하는 경우는 사람이 눈으로 일일이 하는 것 말고, 다른 대안을 찾아야만 했다. 그래서, 이러한 HGP(사람의 염기서열을 밝히고자 하는 프로젝트)가 시작되면서 생물정보학도 같이 세상에 등장하게 되는 것이다. (그림 1)은 미국의 Human Genome Research Institute의 Francis Collins 박사가 제시한 향후 생물학 분야에 있어서의 청사진의 일부분이다. 여기서 계산생물학인 computational biology도 하나의 중요 기둥으로 자리잡고 있는 것을 볼 수 있다[1].



(그림 1) Collins가 제시한 향후 Genomics의 발전 방향에 대한 삽화로 본 논문에서 이야기하고자 하는 생물정보 분석인 Computational Biology가 하나의 큰 기둥으로 자리 잡고 있다.

현재는 HGP도 마무리 되면서 밝혀진 염기서열을 해석하는 분야에 많은 투자와 관심이 집중되고 있다. 염기서열을 해석한다는 것은 30억 베이스 중에서 어느 특정부분이 어떻게 단백질로 되어서, 세포내외에서 특정 역할을 어떻게 수행하는지를 밝히는 것이다. 이러한 해석과정을 연구하기 위한 수많은 연구분야가 있다. 먼저, 염기서열은 A, C, G, T의 4가지 문자로 이루어진 문자열이라고 생각할 수 있는데, 이 문자열의 어느 부분문자열들이 단백질로 될 수 있는 부분인지를 찾는 유전자 예측(gene prediction)이 있을 수 있다. 다음으로 이들 염기서열에서 개인별 차이를 보이는 부분을 SNP라고 부르는데, 이런 SNP 데이터를 이용하여 질병과의 연관성을 분석하는 분야가 있다. 다음으로 최근에 가장 많은 실험을 수행하고 있는 DNA 칩 기술을 이용한 유전자의 기능해석 분야가 있다. 유전자 기능해석에 관심을 두고 있는 연구분야를 특별히 functional genomics라고 한다. 마지막으로 단백질 상태에서 수행한 실험결과 분석 방법으로, 단백질-단백질 상호작용 데이터를 이용한 기능해석과 단백질의 구조를 이용한 기능해석 분야가 있는데 이러한 분야를 일반적으로 proteomics라고 이야기한다. 본 논문에서는 이러한 여러 바이오인포매틱스 분석 분야 중에서 크게 functional genomics와 proteomics로 나누어 기술동향에 대해서 설명을 하도록 한다. 그리고, 여러 바이오 데이터에 대한 정보를 얻을 수 있는 데이터베이스들에 대한 소개를 하고, 마지막으로 이런 바이오인포매틱스 분야의 시장 동향에 대한 것도 시장보고서를 기준으로 설명을 한다.

Ⅱ. 바이오 정보 기술 동향

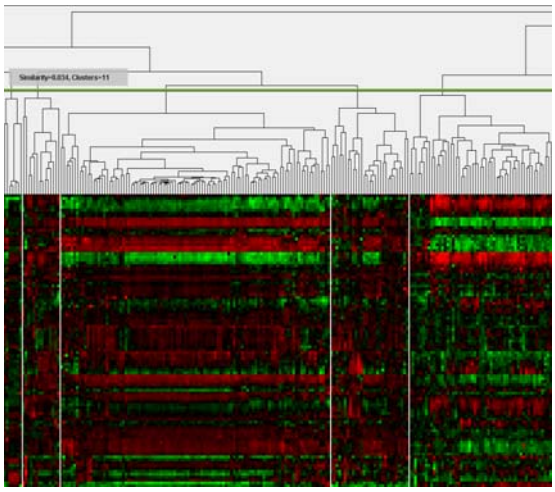
1. Functional Genomics

최근 유전자에 관한 정보를 얻기 위하여 수천, 수만 개의 유전자를 한 번에 실험할

수 있는 마이크로어레이 칩을 많이 사용한다. 이 실험을 통하여 다양한 조건에서 서로 다른 발현 양상을 보이는 유전자 발현데이터가 대량으로 생산되고 있으며, 데이터의 분석과 해석에 매우 관심이 높다.

유전자 발현데이터를 분석하는 가장 기본적인 목적은 기능을 알지 못하고 있는 유전자들을 기능이 알려진 다른 유전자들과 비교하여 그 기능을 예측하기 위함이다. 이때 사용하는 방법이 유사한 발현패턴을 갖는 유전자끼리 묶는 클러스터링이 있다. 클러스터링을 수행하여 기능이 알려지지 않은 유전자들과 발현 패턴이 유사한 기능이 알려진 다른 유전자들의 기능 정보를 가져와서 유전자 해석을 하는 것이다.

클러스터링 방법에는 발현 패턴이 유사한 유전자들을 이웃하는 트리 형태로 구성하는 계층적 클러스터링 방법[2]과 K개의 유사한 발현패턴 그룹인 클러스터로 나누는 군집형 클러스터링 방법이 있다[3],[4]. 계층적 클러스터링인 hierarchical 클러스터링 방법은 (그림 2)와 같이 클러스터링 결과를 트리 모양인 덴드로그램(dendrogram)으로 시각화하여 전체적인 발현패턴을 파악하기는 좋으나, 데이터를 특정 K개의 클러스터로 나누기 어렵다. 군집형 클러스터링 방법인 K-Means나 SOM은 전체 데이터를 분석자가 원하는 K개의 클러스터로 나누지만, 클러스터링 결과가 초기치의 영향을 많이 받는다는 단점이 있다.



(그림 2) Hierarchical 클러스터링의 덴드로그램에서 하위 클러스터를 생성하는 모습

여러 가지 클러스터링 방법을 사용하여 클러스터를 생성한 이후에 그 클러스터가 어떤 의미를 갖는지 또 클러스터가 잘 묶였는지에 대해 해석할 필요가 있다. 클러스터에 속한 유전자들이 공통적으로 갖는 특징을 파악하여, 클러스터를 해석할 수 있으며 클러스터 내의 또는 클러스터 간의 유사도(homogeneity)나 분할도(separation)를 수학적으로 계산하여 클러스터가 잘 묶였는지를 평가할 수 있다[5].

그러나 클러스터에 속한 유전자들이 공통적으로 가지는 특징이라 하더라도, 그 특징이 대부분의 유전자에서 나타나는 것이라면 클러스터의 대표 특징이라 하기 어렵다. 또한 클러스터가 잘 묶였는지 수학적인 계산법으로 평가하는 경우, 클러스터의 묶인 정도를 수치화 함으로써 비교 가능하다는 장점이 있으나, 각 클러스터에 속한 유전자 데이터의 생물학적 의미를 반영하지 못한다. 만일 클러스터의 수학적인 평가척도는 나쁘지만, 같은 기능을 하는 유전자들로 묶여 있다면 분석의 관점에 따라 잘 묶여진 클러스터로 해석할 수 있기 때문이다[6].

이에 대한 방안으로 체계화되어 있는 생물학 온톨로지(ontology)를 이용하여서 클러스터의 특징을 해석하는 방법이 있다. 대표적인 생물학 온톨로지인 GO[7]를 이용하여 클러스터의 대표 특징을 파악하고, 대표 특징의 p-value를 hyper-geometric distribution을 이용하여 제공함으로써 클러스터의 생물학적 특징의 유의미성을 통계학적인 관점에서 평가한다.

클러스터의 해석에 사용하는 GO는 서로 다른 바이오 데이터베이스에 있는 gene product에 대한 일관성 있는 주석정보의 필요에 의해 시작된 프로젝트로서 종(organism)에 독립적이고, biological processes, cellular component와 molecular function의 세 가지 카테고리로 구조화된 생물학 온톨로지이다. 여기에 사용된 GO 용어 간에는 부모-자식의 관

계가 설정되어 있으며, 전체적으로는 DAG 구조로 되어 있다.

GO를 이용하여 클러스터의 대표특징을 파악하는 방법은 먼저 클러스터에 속하는 유전자들을 gene product로 대응시킨 후, gene product와 연결되는 GO 용어의 분포를 조사한다. 따라서 클러스터에 속하는 유전자들이 어떤 GO 용어에 많이 속하는지, 어떤 대표 GO 용어로 요약되는지 파악할 수 있다. 또한 대표 GO 용어가 우연히 뽑힐 확률인 p-value를 다음과 같이 계산함으로써 통계적인 유의미성을 검증해 볼 수 있다.

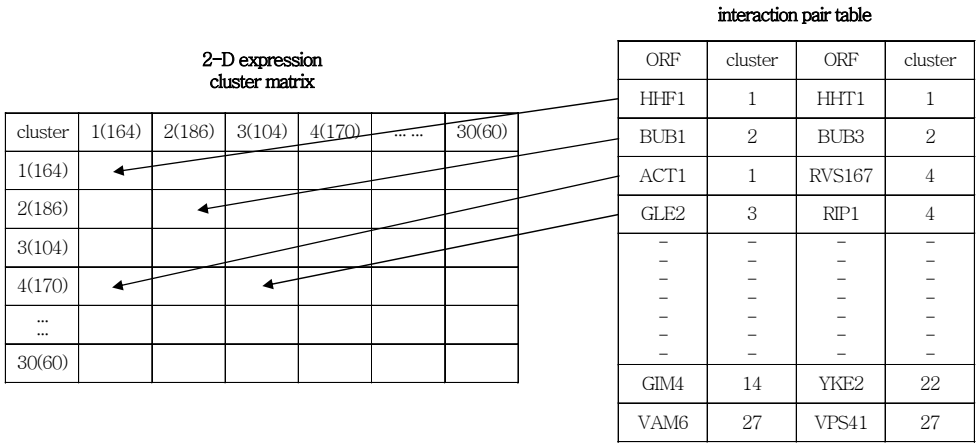
$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

여기서, G는 주어진 종 내에서 전체 유전자의 개수, C는 주어진 GO 용어를 주석정보로 가지는 유전자 개수, n은 클러스터 내 유전자의 개수, 그리고 k는 클러스터 내에서 주어진 GO 용어를 주석정보로 가지는 유전자의 개수를 나타낸다.

위 식의 의미는 n개의 유전자로 이루어진 클러스터 내에서 주어진 GO 용어를 주석으로 가지는 유전자의 개수가 k개 이상인 경우의 확률을 구하는 것이다. 이 확률이 작을수록 우연히 k개의 유전자가 해당 GO 용어를 가지기 어렵다는 뜻이다. 즉, 클러스터를 대표하는 GO 용어의 p-value 값이 작을수록 통계적으로 유의미하다. 클러스터를 대표하는 GO 용어가 GO에서 상위에 있을수록 전체 유전자에서 GO 용어에 속하는 유전자들이 많아지므로 p-value가 낮아지게 된다. 따라서, 클러스터를 대표하는 GO 용어의 p-value 값이 작을수록 클러스터가 잘 묶인 것으로 해석할 수 있다.

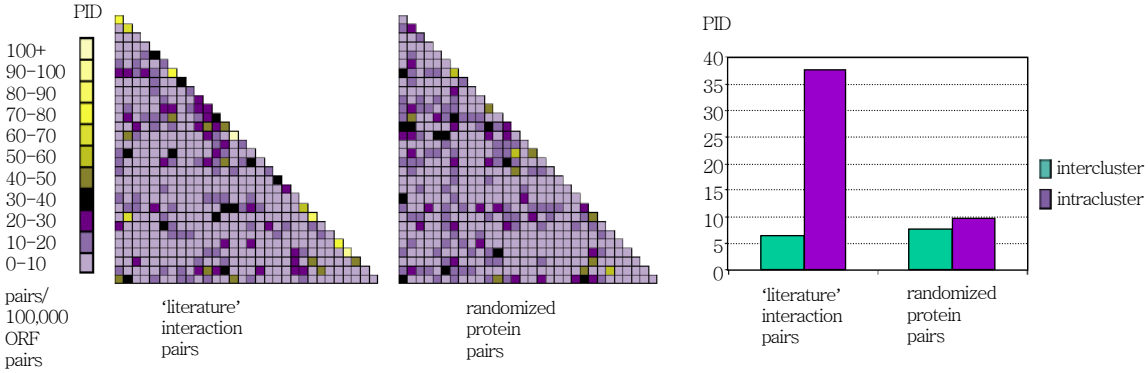
2. Proteomics

단백질 실험 데이터를 이용한 기능 예측 분야에서 가장 많이 사용하는 데이터는 PPI 데이터이다. PPI는 두 개의 단백질들이 상호작용하고 있다. 또는 결합관계에 있다는 것을 의미한다. 이러한 PPI는 생체 내에서 많은 기능을 수행하게 되는데, [8]과 [9] 등의 연구결과에 의하면 유전자들의 발현 양상과 이들 유전자들의 해당 단백질들 간의 상호작용에 상관관계가 있다는 것이 밝혀졌다. Ge 등은 (그림 3)에서처럼 유전자 발현 데이터에 기반한 클러스터링 매트릭스와 단백질 상호작용 테이블을 만들어서 같은 클러스터에 들어 있는 유전자들의 해당 단백질들 사이의 상호작용 빈도가 그렇지 않은 것보다 상당히 높은 것을 확인하였다[8]. (그림 4)에서 보면 "intercluster" 즉 같은 클러스터에 들어가지 않는 단

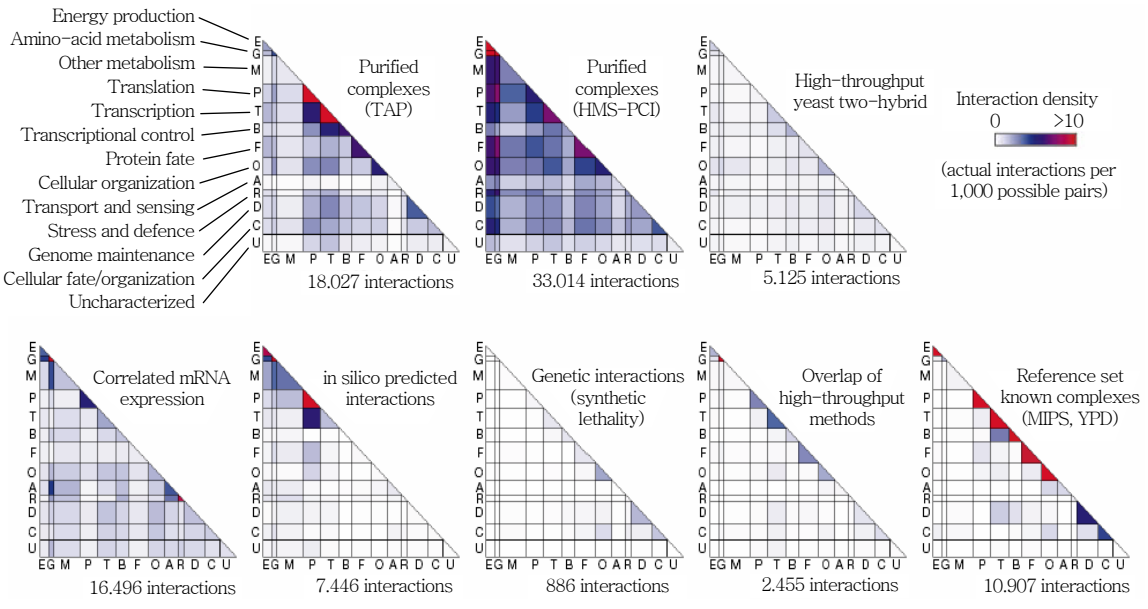


(그림 3) 유전자의 발현 패턴 기반 클러스터링 매트릭스와 각 유전자와의 해당 단백질들간의 상호작용 데이터를 이용하여, 유전자 발현패턴과 단백질 상호작용간의 관계에 대한 분석을 수행하는 방법에 관한 것이다.

백질들의 상호작용이 “intracluster” 즉, 같은 클러스터에 들어간 단백질들의 것보다 그 빈도가 훨씬 작음을 알 수 있다[8]. 이것은 유전자들의 발현 패턴이 유사한 경우에 해당 단백질들이 상호작용할 가능성이 높다는 것을 의미한다. 이것에 앞서 설명한 DNA 칩 실험을 통한 유전자 기능 해석처럼 단백질 상호작용 데이터를 이용한 기능해석이 가능하다는 것을 의미한다고 할 수 있다.



(그림 4) “intercluster”는 같은 클러스터에 속하지 않은 단백질간의 상호작용을, “intracluster”는 같은 클러스터에 속한 단백질간의 상호작용을 의미하는데, “intracluster”의 빈도수가 월등히 많은 것으로 보아, 유전자의 발현 패턴이 유사하면 단백질 간의 상호작용이 있을 가능성이 높음을 알 수 있다. 왼쪽 상호작용 그림에서 왼쪽은 실제 PPI 데이터를 기반으로 한 것이고 오른쪽은 가상의 PPI를 랜덤으로 구성한 것인데, 가상의 PPI에서는 유전자 발현 패턴과의 연관성이 뚜렷이 나타나지 않음을 알 수 있다.



(그림 5) 기능이 알려져 있는 단백질들에 대해서 각 기능들 내에서 상호작용 빈도를 조사한 결과이다. 각 그림들이 의미하는 것은 PPI 실험 방법들에 관한 것으로 첫번째는 TAP라는 실험방법, 세번째는 Y2H 방법에 의한 실험 등이다. 색상은 상호작용 빈도가 높을수록 빨간색을, 낮을수록 흰색을 나타낸다. 세로는 알려져 있는 기능들에 대한 설명이다.

(그림 5)에서는 해석 정보가 있는 단백질들 즉, 단백질의 기능이 알려져 있는 경우에 같은 기능을 가지고 있는 단백질 간의 상호작용 빈도에 대해서 조사를 한 결과이다[8]. 이 결과가 의미하는 것은 PPI상에서 두 개의 단백질이 상호작용하는 경우 이 두 단백질은 유사한 기능을 수행할 확률이 상당히 높다는 것을 의미하고 있다. 따라서, guilt-by-association rule에 기반하여 그 기능이 알려져 있지 않은 단백질을 PPI상에서 상호작용하는 다른 단백질들의 기능을 조사하면 그 단백질의 세포 내 기능과 세포 내 위치 등을 예측할 수 있다는 것을 의미한다.

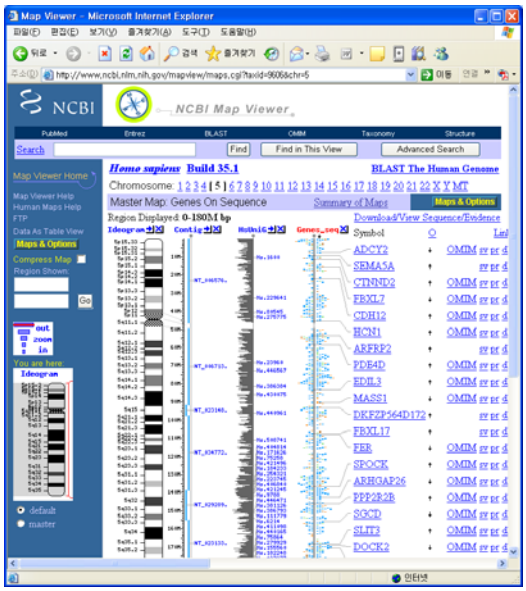
이러한 연구 결과를 바탕으로 하여 단백질의 기능을 예측하고자 하는 많은 연구가 진행되었다 [10]. Schwikowski 등은 기능이 알려져 있지 않은 단백질과 PPI에서 이웃하고 있는 단백질들의 기능들을 리스트상에서 정렬하여 가장 많이 나타나는 기능을 그 단백질에 부여하는 아주 간단한 방법을 사용하였다.

이러한 방식으로 PPI 실험 데이터를 이용하여 다른 정보를 예측하고자 할 때는 한 가지 고려해야 할 것이 있다. PPI 데이터는 그 실험 디자인 자체가 아직까지 신뢰성이 높다라고 말할 수는 없는 수준이기 때문에, 반드시 이용하고자 하는 PPI 데이터에 대한 신뢰도 확보가 우선되어야 한다. 예를 들어, 여러 실험방법들에서 같이 나타나는 즉, A, B 두 개의 단백질은 Y2H 방법을 이용하여도 상호작용이 나타나고, TAP 방법을 이용하여도 상호작용이 나타난다고 하면 높은 점수를 부여하는 방식으로 A-B 상호작용에 대한 신뢰도를 부여하는 방법이라든지, 또는 다른 여러 종에서 같이 나타나는 상호작용에 대해서는 높은 신뢰도를 부여하는 등과 같은 분석이 우선되어야 좋은 결과를 얻을 수 있다.

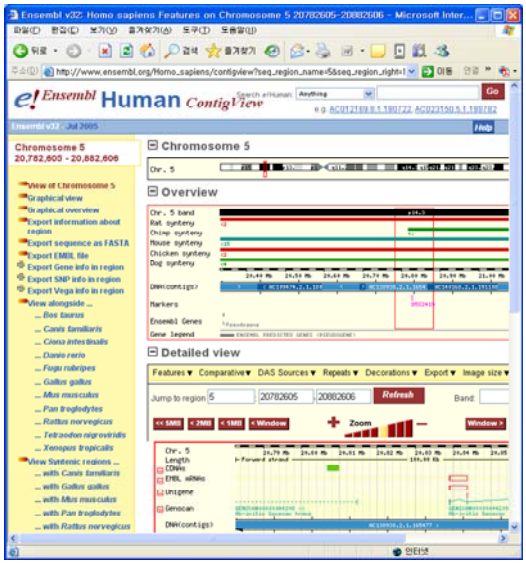
3. Bio-Database Service

앞서 설명에서는 주로 바이오 데이터를 이용한 분석에 대한 설명을 주로 하였다. 이번에는 이러한 바이오 데이터(주로 염기서열 정보, 유전자 해석 정보 등)를 얻을 수 있는 3개의 바이오 데이터베이스에 대해서 소개한다. 이들은 일단적으로 genome browser 형태로 서비스하고 있다. 각 바이오 개체에 대한 정보는 상당히 복잡하게 얽혀 있으며, 하나의 유전자 이름만 알면 관련된 모든 정보를 쉽게 얻을 수 있도록 구성되어 있다.

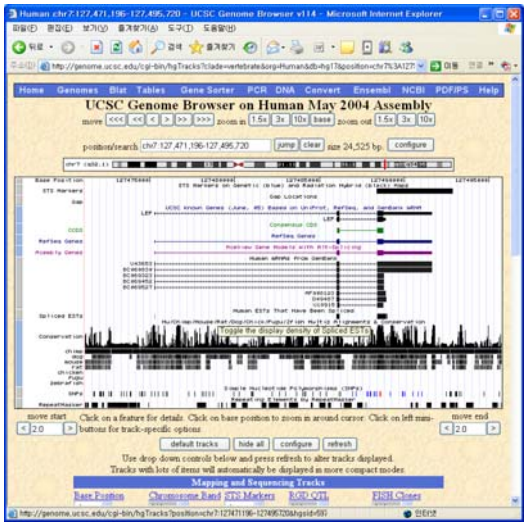
현재 연구자들이 가장 많이 이용하고 있는 데이터베이스는 (그림 6)의 NCBI[11], (그림 7)의 Ensembl[12], 그리고 (그림 8)의 UCSC[13] 등이다. 3개 모두 사용자의 시각적 편의성을 도모하기 위해 genome browser 형태로 서비스를 제공하고 있으며, NCBI와 Ensembl은 생물학자가 설계하였고, UCSC genome browser는 전산학자가 설계하였다는 특징을 가지고 있다. 그래서, 데이터베이스의 설계 효율성이나 시스템 스루풋 등과 같은 것은 다소 UCSC가 앞서고, 시각적인 것은 NCBI와 Ensembl이 앞선 듯한 느낌이다. 이들 데이터베이스들은 FTP



(그림 6) NCBI에서 서비스하고 있는 Genome Browser의 한 모습이다. 각 종별로 염색체 단위로 보여주고 있으며, 유전자 혹은 단백질 이름들과 같은 바이오 개체 정보를 알고 있는 경우에는 "Search" 창을 통해 바로 검색하여 정보를 추출할 수도 있다.



(그림 7) EBI에서 제공하고 있는 Genome Browser인 Ensembl의 모습으로 현재는 사람의 5번 염색체의 일부 모습을 보여주고 있다. Ensembl에서는 자체 아이디뿐만 아니라, NCBI의 바이오 개체 정보도 같이 제공하고 있다.



(그림 8) 미국의 UC Santa Cruz 대학에서 서비스하고 있는 UCSC Genome Browser의 모습이다. 다른 두 개와 다르게 전산학자가 설계하였다는 특징을 가지고 있다. 가로 방향으로 염색체 베이스를 나타내고 세로 방향으로 보고자 하는 각각의 정보들이 놓여 있다. 이들 바이오 정보들은 활성/비활성을 각 사용자가 선택할 수 있도록 구성되어 있다.

서비스를 제공하여 해당하는 모든 바이오 정보를 가져와서 대용량으로 genome 수준에서 분석할 수 있도록 하고 있다.

현재 많은 연구자들이 NCBI 데이터베이스로부터 정보를 얻고 있기 때문에 Ensembl과 UCSC에서는 NCBI 데이터베이스에 바로 접근할 수 있는 링크나 아이디를 같이 제공하고 있는 실정이다.

Ⅲ. 바이오 정보 시장 동향

1. 전 세계 바이오인포매틱스 매출 현황 및 전망

현재까지는 바이오 데이터를 저장 관리하는 서버와 스토리지 등의 하드웨어 시장이 주종을 이루었으나, 데이터 분석 등의 소프트웨어 시장의 증가율이 커지는 추세이다. 이외에도 데이터를 가공한 고부가가치 바이오 콘텐츠 서비스와 콘텐츠 활용 로열티 등의 간접 시장이 있으며 (예를 들면, 바이오 콘텐츠 매출은 2001년 3.3억 달러로서 같은 해 2억 달러의 소프트웨어 매출보다 현저하다[14]), 대부분의 세계적인 제약기업 등 주요 바이오 기업들은 바이오인포매틱스 기술을 내부 역량으로 갖추게 될 것으로 예상되나, 바이오인포매틱스 기술의 특성상 어느 한 기업이 독자적으로 모든 기술을 보유하기는 어려우므로 제약회사와 같은 거대 기업과 소규모 바이오 정보 벤처기업 간의 제휴와 합병이 꾸준히 증가할 것이다.

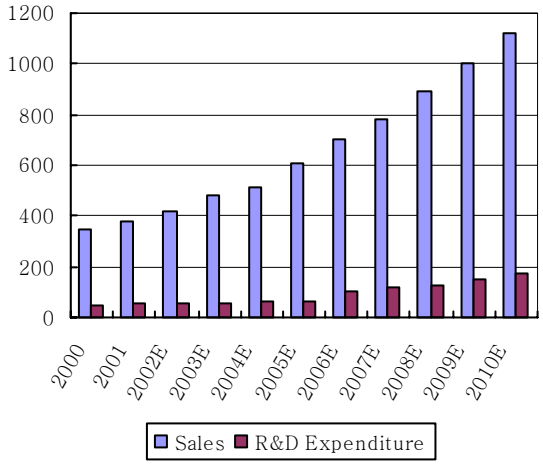
2. 전 세계 바이오인포매틱스 시장

<표 1>의 시장분석 보고서에 따르면, 바이오인포매틱스 시장은 크게 분석 소프트웨어 판매 회사와 enterprise system 판매 업체 및 consulting service provider로 구분된다[15]. 바이오인포매틱스 시장은 대략 2001년에 5억2천만 달러에서 시작하여 2005년에는 20억6천만 달러에 이르고 2010년까지 78억9천만 달러에 달할 것이다. 데이터 저장

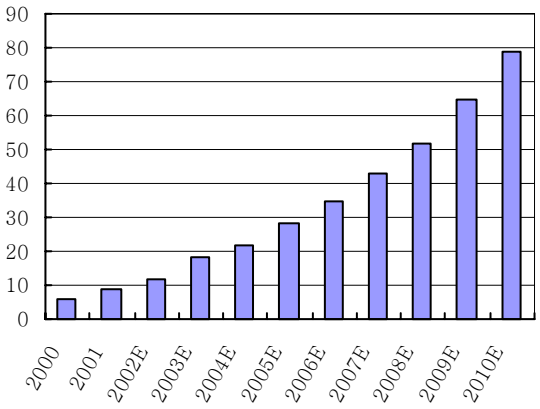
〈표 1〉 바이오인포매틱스 세계 시장

(단위: 백만 달러)

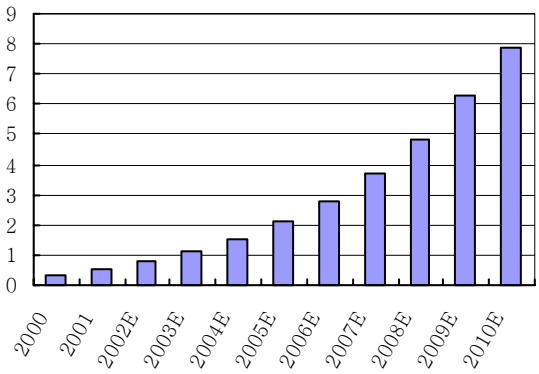
구분/연도	2003	2004	2005	2006	GAGR ('03~'06)
서버/클러스터	4,743.50	5,597.30	6,504.10	7,935.00	17.90
데스크톱	1,399.20	1,501.30	1,603.40	1,747.70	7.40
스토리지	4,978.80	6,572.00	8,583.00	11,844.50	32.30
소프트웨어	3,763.70	4,365.80	5,195.30	6,702.00	19.50
서비스	3,075.60	4,090.60	5,481.40	7,399.80	32.60
네트워킹	674.00	896.40	1,210.10	1,621.50	34.60
기타	156.30	197.00	260.00	356.20	26.60
합계	18,791.00	23,220.40	28,837.30	37,606.80	24.30



(그림 9) 전 세계 제약회사와 바이오테크놀로지 회사의 연구 개발 비용($US Billion)



(그림 10) 전 세계 제약회사와 바이오테크놀로지 회사의 IT 예산($US Billion)



(그림 11) 전 세계 바이오인포매틱스 시장 규모 ($US Billion)

공급자의 현재 시장 크기는 2억6천만 달러이며 분석 소프트웨어 시장은 1억5천만 달러, Enterprise system 공급 시장은 5천만 달러, 공공 데이터베이스 접근 공급자(public database access provider) 및 컨설팅은 각각 3천만 달러 정도이다.

바이오인포매틱스 시장을 정확하게 정의하기는 매우 어렵다. 왜냐하면 바이오인포매틱스 시장이 여러 기술 분야와 응용 분야와 연결되어 있기 때문이다. 또한 바이오인포매틱스와 관련된 산업 분야가 매우 다양하고 대부분의 업체와 관련된 내용이 공개되고 있지 않기 때문이다. 따라서 바이오인포매틱스 시장을 분석하기 위해서는 제약회사와 바이오테크놀로지 회사의 연구 개발 비용 대비 IT 예산을 비교하여 조사를 해 볼 수 있다. 전 세계 제약회사와 바

이오테크놀로지 회사의 연구 개발 비용($US Billion)은 다음과 같다((그림 9), (그림 10) 참조)[15].

전 세계 바이오인포매틱스 시장 규모는 (그림 11)과 같다[15].

3. 세부 분야별 바이오인포매틱스 시장 규모

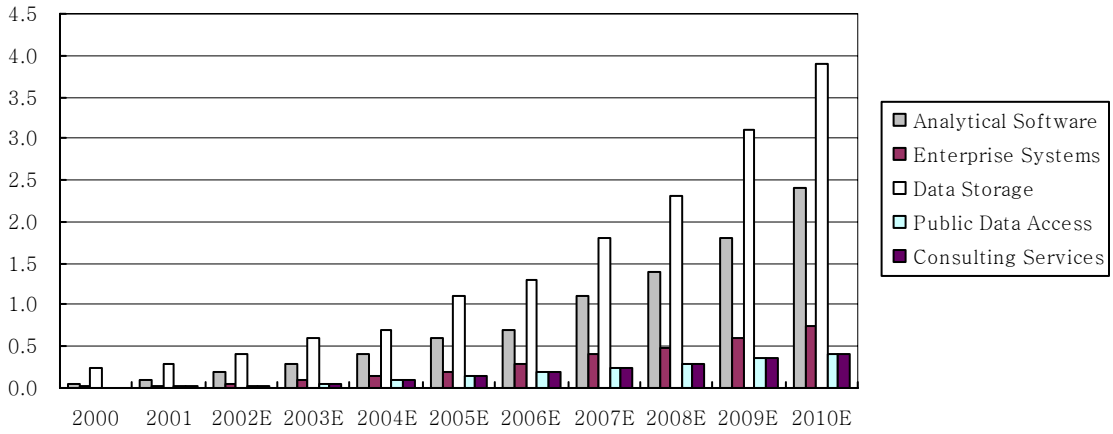
바이오인포매틱스 시장의 커다란 부분은 데이터 저장 공급자가 차지하고 있다. 이는 대략 전체 시장의 50% 정도를 차지한다. 그 다음으로는 분석 소프트웨어 시장으로 30% 정도를 점유하고 Enterprise system 공급자가 10%를, 그리고 공공 데이터베이스 접근 공급자와 컨설팅이 각각 5%씩을 갖는다. (그림 12)는 세부 분야별 전 세계 바이오인포매틱스 시장 규모이다[15].

4. 바이오인포매틱스 분석 소프트웨어 시장

바이오인포매틱스 기술을 이끌 응용 분야는 유전자 서열 분석, 발현 분석, compound library creation, HTS, 단백질 분석, 단백질 구조 결정, PPI, SNP identification and validation, SNP analysis and genotyping 등이 있다. 전 세계의 실험실에서 생성되는 대량의 분자생물/생화학 실험 데이터의 증가와 함께 이러한 데이터를 효율적으로 분석하고 관리하는 소프트웨어에 대한 요구가 증가하고 있다. 이런 각종 바이오 데이터는 각종 소프트웨어 툴에 의하여 분석되고 데이터베이스에 저장되어 공유된다. 바이오인포매틱스 분석 소프트웨어의 범위는 단순 유전자 서열 분석에서부터 PPI까지 그 범위가 넓다. 그리고 최근에는 신약개발을 위한 생물학적 패스웨이의 in silico 모델링을 위한 소프트웨어 개발에 대한 추세가 있다.

생물학적 데이터의 급증은 high-throughput 기기의 발전과 이용에 있다. 이에 따라 넘쳐나는 생물학적 데이터의 관리와 분석을 위한 바이오인포매틱스 도구의 요구가 커지게 되었다. 기존의 genomics 도구에서 점차 산업체들은 proteomics, pharmacogenomics와 chemoinformatics 등의 응용 분야로 그 범위를 넓혀가고 있다. 한편, 바이오인포매틱스 분석 도구 시장의 경쟁도 치열해지고 있다. 따라서 더욱 정교하고 customized된 도구가 늘어나고 있으며 높은 정확도, 경쟁력 있는 가격 및 빠른 분석 속도에 대한 요구가 증가하고 있다. 바이오인포매틱스 분석 소프트웨어 시장은 2004년부터 향후 5년 동안 매년 9.3%의 성장률을 보일 것이다. 이 증가율은 제약업체와 proteomics 분야에서의 분석 도구 채용 증가에 기인한다. (그림 13)은 전 세계 바이오인포매틱스 분석 소프트웨어 시장 전망을 나타낸 도표이다[14].

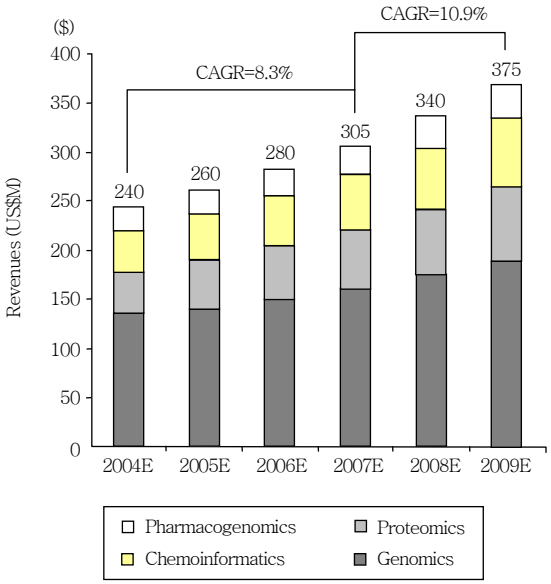
<표 2>에 의하면 학교와 제약업체에서 주력으로



(그림 12) 세부 분야별 전 세계 바이오인포매틱스 시장 규모($US Billion)

사용하고 있는 바이오인포매틱스 분석 도구가 유전자 서열 분석과 유전자 발현 분석 소프트웨어이기 때문에 전체적으로 genomics 분야가 바이오인포매틱스 분석 소프트웨어 시장을 이끌 것으로 전망된다[15]. 하지만 proteomics와 pharmacogenomics 응용분야가 성숙함에 따라 향후 5년 동안 높은 성장률과 시장 점유율을 높일 것이다. 제약업체의 바이오인포매틱스 분석 도구 채용은 현재의 경제 및 경기 동향과 맞물려 서서히 진행될 것이며 이에 따라 분야별 시장 점유율의 변동은 크지 않을 것으로 예상된다. 2004년 바이오인포매틱스 분석 소프트웨어 전체 시장은 2억4천만 달러이며 2009년에는 3억7천5백만 달러로 예상된다. 바이오인포매틱스 분석 소프트웨어 분야별 시장 점유율은 (그림 14)와 같이 예상된다[14].

한편 분야별 성장률은 proteomics와 pharmacogenomics의 증가율이 커질 것이며 genomics의 성장률은 적을 것으로 예상된다.



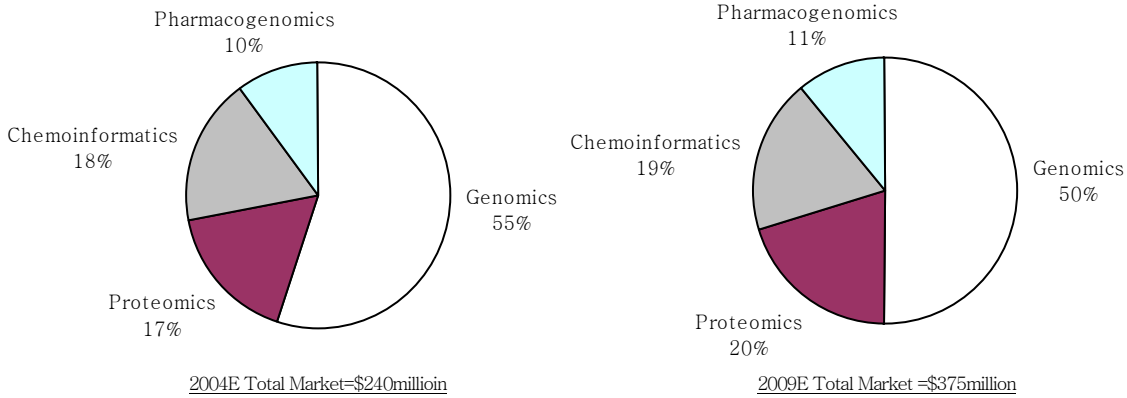
(그림 13) 기술 분야별 바이오인포매틱스 분석 소프트웨어시장 전망

〈표 2〉 바이오인포매틱스 연구 분야별 시장 점유율

Segment	2004년 share	2009년 share
Genomics	55%	50%
Proteomics	17%	20%
Chemoinformatics	18%	19%
Pharmacogenomics	10%	11%

Ⅳ. 결론

바이오 정보 기술과 관련하여 현재 기술 및 시장 동향에 대해서 살펴보았다. 바이오 정보 기술은 IT와 BT의 융합 기술로서, IT 기술력뿐만 아니라, BT에 관한 지식도 가지고 있어야 접근할 수 있는 분야이다. 현재 전 세계는 바이오 정보 기술을 이용하여 BT 분야의 발전에 많은 투자와 노력을 기울이고 있는 상황이다. 우리나라도 2000년경부터 투자하기



(그림 14) 연도별 바이오인포매틱스 분석 소프트웨어 분야별 시장 점유율

시작하였지만, 미국 등과 같은 이 분야 선두주자에 아직도 많이 처져 있는 실정이다. 현재 기술 발전의 큰 장애 요소는 국내외 시장의 미성숙을 들 수 있다. 여러 시장 분석 보고서에 의하면 향후 시장 발전이 가장 기대되는 분야임에도 불구하고 현재 시장 형성은 미미하다. 그래서, 미국과 일본 EU 등에서는 국가적 차원에서 미래 대비하여 투자를 지속적으로 하고 있는 실정이다. 이러한 여러 지표들을 볼 때, 바이오 정보 기술 분야에 대한 국가적으로 지속적인 투자가 필요하다고 생각한다. 그리고, 바이오 정보 분석에 관한 향후 추세는 각 종별로 염기서열이 다 밝혀짐에 따라, 비교유전학(comparative genomics)과 시스템적 분석인 genome-wide analysis 방향으로 발전할 것으로 생각한다.

약어 정리

DAG	Directed Acyclic Graph
DNA	Deoxyribo Nucleic Acid
GO	Gene Ontology
HGP	Human Genome Project
HTS	High Throughput Screening
PPI	Protein-Protein Interaction
SNP	Single Nucleotide Polymorphism
TAP	Tandem Affinity Purification
Y2H	Yeast-2-Hybrid

참 고 문 헌

[1] Francis S. Collins, Eric D. Green, Alan E. Guttmacher, and Mark S. Guyer, "A Vision for the Future of Genomics Research," *Nature*, Vol.422, 2003, pp.835-847.

[2] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-wide Expression Patterns," *PNAS*, Vol.95, 1998, pp.14863-14868.

[3] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting Patterns of Gene Expression with Self-organizing Maps: Methods and Application to Hematopoietic Differentiation," *PNAS*, Vol.96, 1999, pp.2907-2912.

[4] S. Tavazoie, J.D. Hughes, M.J. Campbell, and G.M. Church, "Systematic Determination of Genetic Network Architecture," *Nature Genetics*, Vol.22, 1999, pp.281-285.

[5] J.W. Seo and B. Shneiderman, "Interactively Eploring Hierarchical Clustering Result," *IEEE Computer*, Vol.35, 2002, pp.80-86.

[6] A. Clare and R.D. King, "How Well Do We Understand the Clusters Found in Microarray Data?," *In Silico Biology,* 2002, p.20046.

[7] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock, "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, Vol.25, 2000, pp.25-29.

[8] H. Ge, Z. Liu, G.M. Church, and M. Vidal, "Correlation between Transcriptome and Interactome Mapping Data from Saccharomyces Cerevisiae," *Nature Genetics,* Vol.29, 2001, pp.482-486.

[9] R. Jansen, D. Greenbaum, and M. Gerstein, "Relating Whole-genome Expression Data with Protein-protein Interactions," *Genome Research*, Vol.12, 2002, pp.37-46.

[10] Benno Schwikowski, Peter Uetz, and Stanley Fields, "A Network of Protein-protein Interactions in Yeast," *Nature Biotechnology,* Vol.18, 2000, pp.1257-1261.

[11] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Wolfgang Helmberg, David L. Kenton, Oleg Khovayko, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Joan U. Pontius, Kim D. Pruitt, Gregory D. Schuler, Lynn M. Schriml, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Grigory Starchenko, Tugba O. Suzek, Roman Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko, "Database Resources of the National Center for Biotechnology Information," *Nucleic Acid Research*, Vol.33, 2005, pp.D39-D45.

[12] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X.M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Long-den, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney, "Ensembl 2005," *Nucleic Acid Research*, Vol.33, 2005, pp.D447-D453.

[13] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent, "The UCSC Genome Browser Database," *Nucleic Acid Research*, Vol.31, 2003, pp.51-54.

[14] Navigant Consulting Inc., "Bioinformatics Analytical Software: A Strategic Market Outlook," *Front Line Strategic Market Report*, 2004.

[15] Digital Vector, "Bioinformatics Opportunities," *Digital Vector Market Report*, 2004.