특 집

바이오 지식 생성을 위한 IT 기반 기술

박 선 희
한국전자통신연구원 바이오정보연구팀

Ⅰ. 서 론

21세기 바이오 산업의 중요성을 인식한 세계 각국에서는 세포내의 현상을 다각적으로 이해하기 위하여 다양한 바이오 데이터를 축적하고 이를 바탕으로 고부가가치의 정보를 생성하는 바이오인포매틱스 연구를 활발히 진행시키고 있다.

원시 바이오 데이터는 크게 염색체의 염기 배열에 관한 서열 데이터, 유전자 발현 정보를 지니고 있는 칩 데이터, 단백질의 구조 데이터, 논문이나 특허 등의 문헌에 수록되어 있는 텍스트 데이터 이렇게 4가지로 분류될 수 있다. 이러한 원시 데이터로부터 유전자나 단백질과 같은 바이오 객체들의 각각의 기능과 유기적으로 관련이 되는 총체적인 기능을 밝혀내는 것은 매우 중요한 여러 가지의 의미를 갖는다. 우선 신약개발에서 있어서 향후 바이오인포매틱스가 도입되지 않는 제약회사는 도산할 것이라는 예측이 있을만큼 핵심 역할을 수행하고 있다. 바이오인포매틱스라는 도구를 가지고 가공된 바이오 정보를 통하여 종전의 신약개발에서 수행되었던 수없이 많은 시행착오를 건너뛸 수 있어서 비용과 시간이 절반 이하로 줄어든다는 보고가 있다. 의료산업에서도 유전자 레벨에서의 새로운 진단 방법이 출현하고 있고 이는 획기적으로 진단 및 치료 수준을 높일 것으로 기대되고 있다. 이외에도 유전자 가공 식품, 환경 미생물, 법의학 등 관련 분야는 우리의 거의 모든 일상생활에 걸쳐 있다.

바이오인포매틱스 기술은 1) 대용량 바이오 정보들 간의 관련성을 파악하기 위한 새로운 알고리즘 및 통계학적 처리 방법 연구 2) 유전자나 단백질 등 다양한 바이오 정보의 분석 및 해석 3) 서로 다른 유형의 바이오 정보를 효율적으로 관리하고 사용하기 위한 툴 개발 및 응용 등으로 기술되며 이들은 각각 1) 컴퓨터 사이언스, 통계학, 수학 2) 계산 생물학, 데이터 마이닝 3) 소프트웨어 엔지니어링, 데이터베이스 관리 등의 과학 기술 분야의 지식을 필요로 한다. 이렇게 방대한 바이오인포매틱스 기술 분야 중에서 본 연구에서는 원시 데이터를 가공하여 바이오 정보를 생성하는 바이오 데이터 마이닝 방법과 분산된 바이오 데이터를 효율적으로 통합관리하는 방법에 대해서 알아본다.

Ⅱ. 바이오 데이터 마이닝 및 통합관리 요소 기술

1. 텍스트 마이닝 기술

생물학 관련 분야의 논문이나 지식베이스 등에 있는 텍스트로부터 유용한 정보를 검색 및 추출하는 기술에 대한 연구가 진행되고 있는데 이를 바이오 텍스트 마이닝이라고 한다. 바이오 텍스트 마이닝 기술은 바이오 문헌을 대상으로 자연어 처리 및 정보 추출 기술을 적용하여 생물학적 개체들에 대한 정보와 이러한 개체들 간의 관계를 추출함으로써, 이들의 기능과 상호관계를 유추해내는 기술이다. 현재 바이오 문헌을 대상으로 한 정보 추출 및 가공은 실험적 수준에 머물러 있으며, 효과적인 정보 추출을 위한 다양한 시

도가 계속되고 있다.

바이오 텍스트 마이닝을 위한 중요 요소 기술로는 자연어 처리기술, 정보 추출 기술, 정보 검색기술 등이 있다. 자연어처리 기술은 형태소 해석, 품사 태깅, 구문 분석, 의미 해석, 모호성 해소 등의 세부 기술들로 이루어지고, 정보추출 기술은 텍스트로부터 특정 사실(fact)을 찾아내는 단계와 이러한 사실들이 갖는 관계를 유추해 내는 단계의 크게 두 단계로 이루어진다. 방대한 바이오 정보로부터 보다 빠르게 사용자들이 원하는 정보들을 선택적으로 검색해 사용할 수 있도록 하기 위해 바이오 문헌을 대상으로 자연어 처리 및 정보추출 기술을 적용하고 이에 기본적인 정보 검색 기술들을 결합하여 정련된 데이터를 제공하는 것이 바로 바이오 텍스트 마이닝의 연구 목적이다.

이러한 기술중 가장 기본적인 것이 생물학적 요소들의 개체명, 즉 단백질명이나 유전자명 등을 추출하고 인식하는 기술이다. 이는 단백질, 유전자, 약물, 질병 등의 생물학적 개체들과 관련된 이름들을 개체명 사전, 명명 규칙, 개체명 추정 기법 등을 이용해 인식함으로써, 바이오 문헌 내용의 주체들을 추출하는 기술이다.

개체명 인식을 위한 접근방법은 크게 규칙 기반의 개체명 인식과 통계 기반의 개체명 인식, 그리고 두 가지 방법을 통합한 하이브리드 방식이 있다. 규칙 기반의 방법은 생물학적 개체명 인식을 위한 규칙을 수작업으로 구축하고, 다양한 사전을 이용하여 개체명을 추출하는 방법이다. 규칙 기반의 연구는 가장 높은 성능을 내지만 응용 도메인에 따라 사전 및 규칙이 수동으로 바뀌어야 하므로 비용이 많이 든다는 단점이 있다. 통계에 기반한 방법은 학습 데이터로부터 개체명 인식에 필요한 지식을 자동적으로 학습하는 방법으로 결정트리, 은닉 마르코프 모델 (HMM), 최대엔트로피를 이용한 방법이 대표적이다. 이 방법은 규칙 기반의 방법에 비해 유연성이 있지만, 대량의 학습데이터를 필요로 한다. 하이브리드 방식은 규칙 기반의 방법과 통계 기반의 방법을 통합하여 좀더 나은 성능을 얻기 위한 목적으로 통계 기반의 모델에 규칙이나 어휘 정보, 사전 정보 등의 다양한 지식들을 결합하는 방식이다[1-2].

효율적인 검색을 위해서는 바이오 문헌 정보 분류 및 클러스터링 기술이 중요하며 이는 생물학적 분류체계에 따른 바이오 문헌 자동 분류 기술과 바이오 문헌 집합 클러스터링 기술로 구분된다. 바이오 문헌 자동 분류 기술은 생물학적 분류 정보와 문서 분류 기법을 이용해 바이오 문헌을 생물학적 분류체계에 부합되는 형태로 분류하는 기술이며 바이오 문헌 집합 클러스터링 기술은 사용자가 선택한 특정 바이오 문서 집합을 해당 집합 내에 있는 문서들 간의 응집성에 따라 군집화하여 제시하는 기술이다.

이렇게 추출된 정보를 이용하기 위해서는 이들 간의 관계를 구성하는 것이 중요하며 이를 위하여 바이오 문헌에서의 관계 마이닝 기술이 필요하다. 관계 마이닝 기술은 생물학적 구성요소들 간의 관계 추출 기술을 개발한다. 이를 위하여 생물학적 개체명들을 포함하는 구문들을 문서 내에서 추출해내고 이 구문들에 있는 각 개체명들 간의 관계를 인접단어나 관계기술어, 관계 규칙 등을 이용해 식별함으로써, 실제적인 생물학적 구성요소들 간의 관계를 추출한다. 한편 생물학적 구성요소들 간의 관계들을 유기적으로 결합하기 위한 관계망 구성 및 표현 기술이 필요하며 이것은 바이오 문헌에서 추출된 생물학적 구성요소들 간의 관계들을 관계결합 규칙에 따라 결합함으로써 하나의 망을 구성하고, 이러한 망에서 생물학적 구성요소들의 기능 및 생물학적 경로를 추정 기술이다. 최종적으로 생물학적 구성요소들 간의 관계들에 가중치 부여 기술 개발이 필요하며 이는 추출된 생물학적 구성요소들 간의 관계에 대한 다양한 가중치 부여 기법을 연구함으로써, 각 관계들의 중요도를 표현하는 것이다[3-4].

이러한 바이오 텍스트 마이닝 기술을 이용하여 바이오 문헌 마이닝 기술은 유전자 및 단백질의 위치와 구조 예측, 중요 유전자 패턴 발견 등의 전반적인 생명정보학 연구분야에서 기반 지식을 제공하기 위한 사전 분석 도구 및 축적된 다양한 바이오 정보들을 개념적으로 연결시켜 새로운 정

보를 제공하는 바이오 지식 생성도구의 핵심 기술로 쓰여질 수 있다. 생명과학 연구 결과의 산물인 바이오 문헌을 이용하여 유전자간, 단백질간의 상호 관계, 각종 질병의 증상 및 이들의 관계 등 생체내에서 일어나는 각종 현상을 예측할 수 있는 것이다.

2. 서열 데이터 마이닝 기술

서열 데이터를 분석하여 유전체의 배열을 밝히고 각 부분의 기능들을 판단하는 서열 연구는 여러 분야에 응용될 수 있는 바이오인포매틱스의 가장 기본적인 연구 분야인데 실제 유전자, 단백질 분석에 있어서 여러 분야에 응용될 수 있다. 여기에서는 대표적으로 어셈블리, 모티프 분석, 단일염기변이(SNP) 분석의 세 분야에 대한 설명을 하기로 한다.

조각 서열 어셈블리는 한 번에 밝혀내기 힘든 긴 부분을 판독이 가능한 작은 조각으로 만든 뒤 다시 원래대로 복원해 내는 기술이다[5]. 이 분야는 크게 두 부분에 응용될 수 있다. 한 분야는 EST 분석이며 또 한 분야는 DNA 전체 서열 분석이다. EST(expressed sequence tag)란 간단히 말해서 단백질로 변환되는 유전자의 일부분이라고 할 수 있다[6]. 유전자로부터 잇기(splicing) 과정을 거쳐서 mRNA가 만들어지고 다시 이로부터 단백질이 만들어지게 된다. 따라서 우리는 mRNA를 분석할 필요성이 있는데 mRNA는 매우 불안정한 상태이므로 이를 cDNA라는 이중 나선 구조의 안정된 상태로 만들게 된다. 그런데 cDNA 또한 완벽하게 만들어지기가 매우 어렵고 일부분들만 만들어지게 되는데 이러한 조각들이 바로 EST이다. 따라서 원래의 cDNA 서열을 밝혀내기 위해서는 이렇게 만들어진 EST 조각들을 다시 조립할 필요성이 있는데 이러한 과정을 EST 어셈블이라고 한다. 서열 조각 어셈블은 생물체의 DNA 전체 서열을 완전하게 알아내기 위해서도 이용된다. 셀렐라에서는 인간의 DNA 서열을 밝히기 위해 전체 유전체를 모두 작은 조각으로 만든 후 조립하는 방법을 사용하였다. 두 분야 모두 전산학, 특히 문자열 연구 분야의 다양한 알고리즘이 적용된다. 대표적으로 다중서열정렬[7-8](multiple sequence alignment), 쌍정렬(pairwise alignment), 반복부분 찾기(repeat detection) 등이 이용된다.

단백질 구조연구에서의 모티프는 도메인을 이루는 일부분들을 말하는데 서열 연구에서의 모티프는 특정한 기능을 하게 되는 보존된 부위(conserved region)를 말한다[9]. 이 모티프는 주로 발현조절부위, 즉 유전자의 앞부분에 위치한 여러 특정한 부분들을 찾기 위해 많이 쓰인다. 대표적으로 프로모터 탐색 알고리즘에 많이 쓰이게 되는데 이는 특정한 기능을 하는 유전자의 프로모터에는 특정한 염기서열들이 포함되어 있다는 사실들을 이용한 것이다. 따라서 이러한 부위들에 대한 연구를 위해서 근사문자열매칭 알고리즘[7](어느 정도의 오류를 허용하여 문자열을 찾는 알고리즘), 다중서열정렬 알고리즘[7-8], HMM[9], 신경망 기법[9] 등이 활용된다.

SNP란 개인과 개인간의 DNA에 존재하는 한 염기쌍의 차이를 말하며 보통 1000개의 염기당 한 개씩 존재하는 것으로 알려져 있다. SNP 연구는 질병과 관련된 유전자 발견, 의약 관련 연구에 매우 중요하다. 예를 들어 어떤 특정한 병에 걸린 사람들과 걸리지 않은 사람들의 DNA를 분석하여 특정 부위에 어떤 변이가 있는지를 알아냄으로써 그 부위가 해당 질병과 관련된 유전자라는 사실을 알아낼 수도 있다. 또한 어떤 부위의 염기가 특정한 사람들은 특정 약물에 대한 반응이 다른 사람들과 다름을 이용하여 개개인에 대한 최적의 약물을 투여할 수도 있다. 이렇게 한 염기가 바뀜으로 인해(이 부분이 유전자라면) 생성되는 단백질은 전혀 다른 것이 될 수도 있는 것이다. 이러한 SNP를 찾기 위한 접근 방법으로는 이미 매핑된 STS를 이용하는 방법, EST DB를 이용하는 방법 등 여러 가지 방법이 있다.

3. 발현 데이터 마이닝 기술

1996년 미국 Standford 대학에서 DNA chip 기술이 개발된 이후로 수천-수만개의 유전자 발현 실험이 한번에 가능하게 되었고, 다양한 환경

하에서 수많은 유전자들의 발현 양상에 대한 데이터를 얻을 수 있게 되었다. 그러나, 문제는 이렇게 얻어진 대규모의 칩 발현 데이터를 어떻게 효율적으로 분석하여 유용한 생물학적 정보를 얻을 수 있는가 하는 것이다. 현재 일반적으로 칩 발현 데이터를 활용하는 분야는 크게 세 가지로 나누어 볼 수 있다.

첫째는 하나의 칩 위에 심어진 수천 또는 수만 개의 유전자 집합에 대해 여러 다른 조건 하에서의 발현 양상을 측정한 후, 각 유전자의 측정된 발현 프로파일을 기반으로 유사한 발현 양상을 보이는 유전자 그룹을 찾아냄으로써 유사한 기능을 지닌 유전자들을 알아내거나 기능이 밝혀지지 않은 유전자의 기능을 유추하는 데 활용하는 것이다. 이를 위하여 DNA 칩 영상을 처리하여 각 유전자의 발현 정도를 수치화하기 위한 칩 이미지 처리 기술, 품질 제어 등이 필요하며, 수치화된 유전자 발현 프로파일을 분석하여 유사한 유전자 그룹을 찾아내기 위해서는 hierarchical clustering, k-means, self-organizing map, principal component analysis 등의 다양한 클러스터링 기술이 사용되고 있다[10-11]. 또한, 클러스터 결과의 검증을 위해 figure-of-merit, gap-statistic, 생물학적 지식 등이 사용되기도 한다[12].

둘째는 하나의 세포 내의 여러 유전자들 중에서 다양한 조건이나 환경 하에서 특정 질병이나 조직에 특이적인 발현 양상을 보이는 유전자들을 찾아내거나, 미래의 특정 질병에 대한 발생 가능성을 예측할 수 있도록 현재까지 축적된 유전자들의 발현 프로파일과 특정 질병의 발생 여부 또는 진행 정도 등과의 관계를 모델링하는 데에 활용하는 것이다. 이를 위해 유전자 발현 양상을 Euclidean distance, correlation coefficient, cosine coefficient 등의 다양한 특징들로 표현할 수 있으며[13], 이렇게 표현된 특징들에 k-nearest neighbor, neural networks, support vector machine 등과 같은 다양한 분류(classification) 방법을 적용함으로써 유전자 발현 양상과 특정 질병 여부 등과의 연계성을 모델링할 수 있다[14-15]. 생성된 모델의 결과 검증은 대개 cross-validation 방법을 사용하고 있다.

셋째는 칩 위에 심어진 유전자들 간의 상호 조절 관계를 밝혀내는 것이다. 하나의 유전자는 다른 유전자의 발현 양상에 영향을 줄 수가 있는데, 이러한 여러 유전자들 간의 조절 관계를 나타내는 것을 유전자 조절 네트워크라 한다. 이러한 유전자 조절 네트워크를 밝혀내기 위해서는 동일한 유전자 집합에 대해 여러 시간대별(time series)에 실험한 데이터나, 또는 특정 유전자를 제외(knock-out)시켰을 때와 그렇지 않을 때의 유전자 발현 양상에 대한 데이터가 필요하게 된다. 주어진 데이터부터 유전자 조절 관계를 밝혀내기 위해 현재 많이 연구되고 있는 유전자 조절 네트워크 구성 방법은 Boolean network, Bayesian network, differential equations 등이 있다[16-17]. 그러나, 이들 방법은 공통적으로 네트워크 구성을 위한 최적의 해를 찾기 위한 시간 복잡도가 높아, 최근의 연구는 이러한 계산의 복잡도를 줄이면서 가장 알맞은 해를 구하기 위한 방법을 찾는 데 집중되고 있다.

4. 단백질 상호작용 네트워크 기술

유전자의 정보가 발현되어 최종적으로 생성되는 물질로서 단백질들은 상호 유기적인 작용을 통해 세포 내에서 신호전달(Signal Transduction), 세포 주기(Cell Cycle), 분화(Cell Development), DNA 복제(Replication), 전사(Transcription), 번역(Translation), 물질대사(Metabolism) 등과 같은 세포 생리 활성 반응을 조절하게 된다. 따라서, 특정 단백질과 상호작용을 갖는 다른 단백질을 하나의 관계로 나타낸다면 세포 내에서 특정 반응을 조절하기 위해 참여하는 단백질들의 사이의 전체적인 상호작용들은 하나의 네트워크 형태로 표현될 수 있을 것이다[18-20].

이 단백질 상호작용 네트워크는 생물학적인 관점에서 단백질의 기능을 예측하기 위한 중요한 정보로 이용된다. 즉, 하나의 단백질의 기능은 이 단백질과 상호작용을 하는 단백질의 기능과 유사

하다는 기본 가정에 따라 기존에 기능이 잘 알려진 단백질에 대한 정보를 통해 이 단백질과 상호작용을 하는 여러 미지의 단백질들에 대한 기능을 예측할 수 있을 것이다.

단백질 상호작용 관계 정보는 기본적으로 yeast two hybrid 실험을 통해 추출된다. 실험을 통해 구축된 단백질 상호작용 관계 정보는 데이터베이스에 체계적으로 저장되며 관리된다. 대표적인 데이터베이스로 YPD(Yeast Proteome Database), CYPG, PIMdb(Drosophila Protein Interaction Map database), BIND(Biological Interaction Network Database), DIP (Database of Interacting Protein), Path-Calling, Interact, PimRider 등이 있다.

단백질 상호작용 네트워크를 시각화하는 최적화 알고리즘은 중요한 연구분야로 인식되고 있다. 그 이유는 하나의 네트워크가 방대한 노드와 에지로 구성된 그래프 형태로 표현되기 때문에 이들을 시각화하기 위해 엄청난 컴퓨팅 타임을 요구하게 된다. 일반적으로 단백질 상호작용 네트워크 시각화 최적 알고리즘은 NP-hard 문제로 알려져 있으며, 많은 휴리스틱 알고리즘이 개발되어 있다. 대표적으로 Fruchterman & Reingold가 제안한 Force-Directed Placement 알고리즘이 많이 이용되고 있다.

Force-Directed Placement 알고리즘을 예로들면 이 알고리즘은 전역작용(global force), 지역작용(local force) 그리고 좌표 조정(reposition) 과정을 반복하게 된다. 즉, 하나의 노드와 에지를 통해 직접 이웃하는 노드들은 지역작용에 의해 점점 가까워지며, 다른 모든 노드들과는 전역작용에 의해 점점 멀어지는 좌표점을 가지게 된다. 이 반복은 각 노드의 변화량이 특정 임계값 이하가 될 때까지 계속되기 때문에 많은 노드와 에지를 가지는 단백질 상호작용 네트워크에 적용하였을 경우 상당한 계산 시간을 요구하게 된다. 이렇듯 네트워크 모델링 기술의 관건은 사용자가 빠른 시간에 원하는 정보를 얻을 수 있는 최적화된 네트워크를 제공함에 있다.

5. 데이터 통합관리 기술

바이오 데이터베이스 통합에 대한 시도는 1980년대 후반 개별 생물학자들이 발견한 여러 생물의 DNA 정보를 수록하는 공공의 데이터베이스가 만들어지면서부터 라고 할 수 있다. 일찍이, 선진국들은 방대한 양의 지놈 및 유전자 정보를 체계적으로 수정/관리하는 국가 차원의 센터를 두고, 관련 기술을 발전시켜 왔다.

현재, 각종 생물학 분야에서는 빠른 속도로 증가하는 다양한 정보들을 서로 공유하려는 공감대가 형성되어 있는 추세이다. 하지만, 의약 개발, 유전학 등의 각 분야에서는 자신들만의 필요성에 따라 고유한 데이터베이스를 구축해 왔으며, 또한 아주 특성화된 상호 작용과 분석 도구들이 이들을 바탕으로 구축/발전되어 왔다. 그 결과, 이들 정보 서비스들은 서로 분산되어 있고, 이질의 데이터 형태를 가지며, 서로 낮은 의미로 연관되어 있다. 이러한 상황에서, 생물학자가 자신의 데이터 분석 중에 공개 데이터베이스 소스를 대상으로 복잡한 검색 타스크를 수행하기 위해서는 다음의 절차를 거쳐야 한다. 먼저, 1) 생물학자는 각각의 데이터베이스 소스들에 대한 메타데이터 뷰를 가져야 하며, 이 소스들간의 의미적 연관성을 파악하여야 한다(예를 들어, 단백질에 대한 일반 정보는 SWISS-PROT, 단백질 서열상의 motif에 대해서는 PROSITE, 그리고, 서열 일치에 대한 도구는 BLAST식으로). 2) 각기 다른 소스들로부터 데이터 추출에 필요한 각기 다른 형식과 요구에 따라 각 소스별 질의를 작성하여야 한다. 또한, 3) 각각의 소스와 통신하여 결과를 추출하고, 이에 대한 처리를 위해 중간형태의 단일 형태(예, XML)로 변환하여야 한다. 4) 이와 같이, 해당 소스에 적용한 질의들을 계획(plan)하고, 추출된 데이터들을 추적하며 연결시키는 과정을 반복적으로 수행한다. 물론, 각 질의는 내부적으로 데이터를 검색하고, 필터링하며, 요구된 처리를 수행하는 일련의 과정을 거쳐야 한다. 특히, 이러한 과정을 거치기 위해서는 여러 선택적인 방법들이 가능하며, 이를 위한 다양한 옵션들 중에서 선택되어져야 한다.

이러한 복잡한 절차들은 생물학자 자신의 고유 분석 작업에 비해 매우 힘든 작업이다. 특히, 가능한 많은 공개 데이터베이스 소스를 통해 자신의 실험을 검증하려는 노력에 비례하여 그 복잡도는 더욱 증가된다. 따라서, 이러한 과정을 보다 자동화하고 단순화하려는 시도들로, 생물학자가 요구하는 소스의 선택, 조합, 상호 연산하도록 설계된 여러 방법들이 제안되었다. 가장 대표적인 정보 통합 방법들로, 웹 기반 통합 검색에서 데이터 웨어하우스 등의 여러 방식들이 제안되었다[21]. 웹 기반 통합의 한 방법인 SRS(Sequence Retrieval System)[DD]은 DNA/아미노산 서열, 단백질 모티프, 단백질 구조, 그리고 문헌 데이터베이스로부터 각 데이터를 검색/연결, 접근하기 위한 방식을 제공하고 있다. 또 다른 방법으로 Entrez[EE] 역시 비슷하게 DNA/아미노산 서열, 지놈 맵 데이터, 3D 구조, PubMed 문헌 데이터베이스에 대한 접근을 제공한다. 데이터웨어하우스 기술을 이용한 물리적 통합은 성능면에서 장점이 있지만, 스토리지 비용의 증가, 원 데이터 소스에 대한 다양한 분석 기능 적용의 어려움, 계속 변경되는 데이터 소스 변경에 대한 즉각적인 반영의 어려움, 그리고, 새로운 분석 서비스를 위해서는 기존의 모델과 구조가 변경되어야 한다는 큰 제약이 있다.

최근에는 이러한 제약을 극복하기 위한 방법으로, 생물학자가 바라보는 개념적인 모델과 실제 데이터베이스 소스가 제공하는 물리적인 모델을 구분하고, 이들간의 문법적/의미적인 차이를 해소하는 사상 모델로 이루어지는 3단계 모델 구조가 각광을 받고 있으며, BioKleisli[22], TAMBIS[23] 등에 적용되고 있다. 물리적인 모델은 데이터베이스 소스의 주요 데이터 구조와 파일 형식을 정의하고 있으며, 다른 소스들과의 형 일치를 위한 작업을 수행한다. 즉, 이 단계는 각 소스에 대해, 실제 데이터에 대한 접근 경로, 변환에 필요한 다양한 프로그램 인터페이스 모임을 가지는 것으로 위치와 형식에 대한 투명성을 제공하고 있다. 개념적인 모델은 등록된 데이터베이스 소스들에 대해 소스 독립적인 단일화된 개념 단계의 모델 표현과 단일의 질의어를 제공하는 것으로, 생물학자에게 각 소스의 실제 세부 내용을 감춘다. 즉, 이 단계는 각 소스의 개념들을 연관시키고, 소스간의 상호 연산에 대한 기반 프레임을 제공한다. 사상 모델은 질의 처리기 역할로서, 개념 단계의 관점에서 기술된 질의를 실제 물리 단계의 데이터 소스에서 실행되어질 수 있는 요구로 변환하는 역할을 수행한다. 즉, 생물학자가 명세한 질의를 실제 물리 계층의 데이터 서비스와 연관시킨다.

이러한 구조의 장점은 1) 생물학자에게 개념적인 모델을 통해 하나의 데이터 모델과 하나의 질의어를 사용하도록 함으로써, 다수 데이터베이스 소스들에 대해 고수준의 투명성(transparency)을 보장할 수 있다는 것이다. 즉, 각 데이터 소스의 어떠한 지식과도 독립적으로 질의를 표현할 수 있다는 것이며, 2) 매핑 모델을 통해, 요구된 각각의 데이터베이스 소스들에 대해 질의 생성이 가능하게 되며, 또한, 질의 실행 계획과 최적화에도 유리하다. 또한, 3) 다양한 데이터베이스 소스들간의 문법적 뿐만 아니라 의미적인 이질성을 인식하고 파악할 수 있다. 마지막으로, 4) 사상 모델과 물리적 모델의 분리를 통해, 새로운 데이터베이스 소스의 확장 및 자체의 각종 실험 데이터 모델과의 통합에도 유리하다는 장점이 있다. 물론, 데이터웨어하우스 방식과 달리, 데이터를 자체적으로 중복하여 갖지 않으므로, 이들에 대한 통합화 된 뷰에 대한 어떠한 변경도 가능하지 않는 것으로, 오로지 읽기 전용의 접근만이 가능하다.

현재, 데이터베이스 소스에 대한 통합은 실험실 혹은 연구센터에서 도출된 데이터를 검증하고 인증받기 위한 것에서부터, 신약 개발 등을 위해 각 분야에서 누적된 다양한 정보를 도출/예측하기 위한 것과 같이 광범위하게 요구되고 있는 실정이다. 이를 위한, 통합 시스템은 각 데이터베이스 소스에 대한 다양성과 지속적인 확장성, 검색 성능의 보장, 생물학자의 경험 지식 표현 등의 수용이 가장 중요하다고 할 수 있다.

Ⅲ. 결 론

이상과 같이 주요 IT 기반 바이오인포매틱스 기술 개발에 대하여 알아보았다. 차세대 산업을 이끌 바이오 산업을 활성화하기 위하여 최적화된 연구 결과를 도출해 낼 수 있는 방법은 국내의 강한 IT를 이용해야 한다. 미국과 같은 선진국의 경우 BT 관련 기초 연구에 장기간의 시간과 연구비를 투자하여 그 결과 많은 바이오 정보가 축적되어 있으나 우리나라의 경우 최근에서야 연구 개발이 시작되었으며 존재하는 바이오 정보도 산발적으로 펼쳐져 있어 선진국에서 진행되는 바이오인포매틱스와 이에 관련된 모든 분야의 연구 개발은 불가능하다. 우리나라의 IT 기술 및 인프라는 세계적인 기반을 보유하고 있으므로 이를 활용하여 고부가가치의 정보를 추출해 낼 수 있다면 바이오 산업 뿐 아니라 침체 상태에 접어든 IT 산업에도 새로운 산업 분야를 열 것으로 기대된다.

참 고 문 헌

〔1〕 Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T., "Toward IE : Identifying protein names from biological papers", Proceedings of the Pacific Symposium on Biocomputing (PSB 98).

〔2〕 Denys Proux, Francois Rechenmann, Laurent Julliard, "Detecting Gene Symbols and Names in Biological Texts : A First Step toward Pertinent Information Extraction", Genome Informatics, 9, pp.72-80, 1998.

〔3〕 M.Stephens, M.Palakal, S.Mukhopadhyay, R.Raje, "Detecting Gene Relations From Medline Abstracts", Proceedings of the Pacific Symposium on Biocomputing, 2001.

〔4〕 Thomas C. Rindflesch, Lorraine Tanabe, John N. Weinstein, Lawrence Hunter. "EDGAR : Extraction of Drugs, Genes And Relations from the Biomedical Literature", Proceedings of the Pacific Symposium on Biocomputing, 2000.

〔5〕 S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J.P. Mesirov, E.S. Lander, ARACHNE : A Whole-Genome Shotgun Assembler, Genome Research, 12, pp. 177-189, 2002.

〔6〕 T.A. Brown, "Genomes", John Wiley & Sons (ASIA) PTE LTD.

〔7〕 D. Gusfield, "Algorithms on Strings, Trees, and Sequences", Cambridge University Press.

〔8〕 M.S. Waterman, "Multiple sequence alignment by consensus" Nucleic. Acids. Res. 14, pp.9095-9102, 1986.

〔9〕 D.W. Mount, "Bioinformatics : Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press.

〔10〕 P. Tamayo, "Interpreting patterns of gene expression with self-organizing map : methods and application to hematopoietic differentiation", Proc. Natl. Acad. Sci. USA, 96, pp.2907-2912, 1999.

〔11〕 M.B. Eisen, P.T. Spellman, P. O. Brown and D. Botstein, "cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci. USA, 1998

〔12〕 K.Y. Yeung, D.R. Haynor and W. L. Ruzzl, "validating clustering for gene expression data. BioinformaticsT. R Golub, molecular classification of cancer : class discovery and class

prediction by gene expression monitoring", Science, 286, pp.531-537, 1999.
[13] S. Cho and J. Ryu, "classifying gene expression data of cancer using classifier ensemble with mutually exclusive features", Proceedings of the IEEE, 90, November 2002 Issue.
[14] T.S. Furey, N. Cristianini, N.Duffy, D. W. Bednarski, M.Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics", 16, pp. 906-914, 2000.
[15] I. Shmulevich, E.R. Dougherty, S. Kim and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks", Bioinformatics, 18 pp.261-274, 2002
[16] N. Friedman, M. Limial, I. Nachman and D. Peer, "Using Bayesian networks to analyze expression data", Journal of computational biology, 7, pp.601-620, 2000.
[17] Z. Sxallasi and R. Somogyi, "Genetic network analysis", PSB2001 Tutorial.
[18] Uetz, P., Ideker, T. and Schwikowski, B., "Visualization and integration of protein-protein interactions," Cold Spring Harbor Laboratory Press, pp. 623-646, 2002.
[19] Tucker, C. L., Gera, J. F. and Uetz, P., "Towards an understanding of complex protein interaction maps," Trends in Cell Biology, 11, pp.102-106, 2001.
[20] Schwikowski, B., Uetz, P. and Fields, S., "A Network of Protein-Protein Interactions in Yeast," Nature Biotechnology, 18, pp.1257-1261, 2000.
[21] O.Ritter, P.Kocab, M.Senger, D.Wold, and S.Suhai, "Prototype Implementation of the Integrated Genomic Database, Computer and Biomedical Research", 27, pp.97-115, 1994
[22] S.Davidson, C.Oberton, V.Tannen, L. Wong, BioKleisli "A Digital Library for Biomedical Researchs", Intl Journal of Digital Libraries, 1, pp.36-53, 1997.
[23] C.A.Globe, N.W.Paton, R.Stevens, P. G.Baker, G. na, M.Peim, S.Bechhofer, and A.Brass, "Transparent Access to Multiple Bioinformatics Information Source", IBM System Journal, 40, no. 2 Issue, 2001.

저 자 소 개

박 선 희

1981년 2월 서울대학교 사범대학 수학과 졸업(BS), 1986년 8월 Univ. of Texas at Austin 수학과 졸업(MS), 1989년 12월 동대학원 물리학과 졸업 (Ph. d.), 1990년 1월~1994년 6월 : Univ. of Texas at Austin, I.C.T.P. in Italy, 서울대 이론물리센터 등에서 Postdoc, 1994년 7월~현재 : 한국전자통신연구원, 〈주관심 분야 : 바이오인포매틱스, 생체신호처리, 바이오센서〉