정보통신부 BIT관련 연구개발사업 현황/계획

한국전자통신연구원 박선희

1. 서 론

바이오인포매틱스는 IT의 발달사와 생물학 데이터의 발달사와 함께 생물학의 발전을 지원하기 위한 기술로서 진화되어왔다. 포트란의 출현 이후 수많은 프로그래밍 기술, 데이터 마이닝 기술, 분석 기술, DB 기술 등이 여러 종류의 바이오 데이터에 적용되어 생물학자들에게 도움을 줄 수 있는 다양한 소프트웨어가 개발되고 있다.

우리나라는 현재 세계적인 IT인프라가 구축되어 있고 우수한 IT 인력을 보유하고 있다. 또한 BT 측면에서도 최근 줄기세포 등과 같은 세계적인 연구 성과로서 기술을 선도할 수 있다는 자신감을 가지게 되었다. 이러한 배경을 가지고 **정보통신부에서는 BT 산업에 성장 엔진을 달아줄 IT 접목 기술 개발에 지원하고 있고 현재 한국전자통신연구원에서는 정보통신부의 지원 하에 IT 기반 바이오인포매틱스 연구를 수행하고 있다.**

본 글에서는 한국전자통신연구원에서 진행 중인 기술을 중심으로 IT의 관점에서 바이오인포매틱스 기술을 소개하고자 한다.

2. IT기반 바이오인포매틱스기술 개발현황

2.1 유전정보 분석 기술

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 발현 데이터가 대량으로 생산되고 있으며, 데이터의 분석과 해석에 매우 관심이 높다.

유전자 발현데이터를 분석하는 가장 기본적인 방법으로는 유사한 발현패턴을 갖는 유전자끼리 묶는 클러스터링이 있다. 클러스터링 방법에는 발현 데이터가 유사한 유전자들을 이웃하는 트리 형태로 구성하는 계층적 클러스터링 방법과 K개의 유사한 발현 패턴 그룹인 클러스터로 나누는 군집형 클러스터링 방법이 있다. 계층적 클러스터링 방법은 클러스터링 결과를 트리 모양인 댄드로그램으로 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. 그리고 여러 클러스터링 방법을 적용하여 생성한 클러스터들을 해석하는 것도 매우 중요하다. 일차적인 해석 단계로 클러스터에 속하는 유전자들의 공통적인 특징을 파악할 수 있는데, 최근 생물학적 온톨로지인 Gene Ontology[1]나 MIPS를 이용한 방법[2]이 있다.

본 시스템(*GEDA-C*)에서는 기존 클러스터링의 결과를 응용하여 클러스터링하는 seed 클러스터링 방법을 제안하였다. 그리고 생물학적 온톨로지인 Gene Ontology를 이용하여 seed 클러스터링과 K-means 클러스터링 방법으로 생성한 클러스터들을 해석하고 비교해보았다.

2.1.1 Seed 클러스터링

Seed 클러스터링은 발현 데이터가 매우 유사한 유전자들은 여러 가지 클러스터링 방법에서 같이 묶여서 나타나는 특징에 기초하여 제안한 알고리즘이다. 즉 클러스터링 알고리즘이나 파라메터에 민감하지 않으며, 안정적으로 같은 클러스터에 나타나는 유전자들의 집합을 조사하여 이를 클러스터링에 이용하자는 것이다. 안정적으로 같은 클러스터에 나타나는 유전자 집합들은 그 자체로도 충분히 의미가 있으며, 이 유전자들의 발현 데이터를 적절히 이용하여 군집형 클러스터링의 초기치로 이용하면 초기치에 민감한 군집형 클러스터링의 단점을 보완할 수 있다.

Seed 클러스터링은 다음의 세 단계로 이루어진다. 첫 번째, 잘 알려진 클러스터링 방법을 사용하여 클러스터링 한다. 이때 다양한 파라미터를 적용함으로써 여러 가지 클러스터링 결과를 얻을 수 있다. 두 번째, 여러 가지 클러스터링 결과 중에서 같은 클러스터에 같이 나타나는

유전자 집합들을 추출하여 seed를 생성한다. 세 번째, 두 번째 단계에서 추출한 seed를 클러스터링 방법의 초기치로 설정하고 클러스터링 한다. 각 단계별 구현과 알고리즘은 다음과 같다.

① 기존의 클러스터링 방법을 이용: 먼저 알려져 있는 클러스터링 방법을 사용하여 클러스터링 한다. 계층적 클러스터링 방법으로는 Hierarchical 클러스터링을 사용하였다. Hierarchical 클러스터링은 특정 K개의 하위 클러스터를 생성하지 않는다. 이를 보완하기 위하여 그림 1에서처럼 Hierarchical 클러스터링의 결과를 시각화하는 댄드로그램에 세 가지 편리한 인터페이스 기능을 추가하였다. 댄드로그램에서 노드들이 연결된 유사도 값을 보고 적절한 값에서 트리를 절단할 수 있는 절단 기능[3]과 이 절단에 의해 생성되는 하위 클러스터 발현 패턴의 시각화 기능, 그리고 절단하여 생성되는 하위 트리들 중에서 원하는 부분 또는 전체를 하위 클러스터 데이터 집합으로 등록할 수 있는 클러스터 등록 기능을 제공한다. 이를 이용하여 분석자는 댄드로그램에서 절단선을 조정, 생성되는 하위 클러스터들의 패턴을 보고 원하는 하위 클러스터들을 생성할 수 있다.

군집형 클러스터링 방법인 K-means이나 SOM의 경우에는 초기 파라메타 입력 값을 지정함으로써 분석자가 원하는 개수만큼의 하위 클러스터를 생성할 수 있으므로 기존의 알고리즘을 그대로 사용한다.

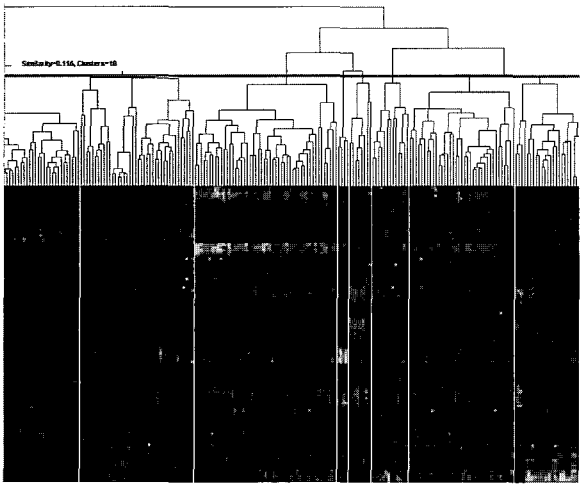

그림 1 Hierarchical 클러스터링의 댄드로그램에서 하위 클러스터 생성

② 하위 클러스터에서 공통으로 나타나는 유전자들을 조사: ①에서 생성한 하위 클러스터에서 같이 나타나는 유전자들을 파악한다. 예를 들어 그림 2에서 A클러스터링 결과와 B클러스터링 결과를 비교해보면 유전자 g1, g2, g8은 같은 클러스터 내에서 같이 나타난다. g6, g7도 마찬가지이다. 이렇게 같은 클러스터 내에서 같이 나타나는 유전자들의 발현 데이터의 평균치를 계산하여 새로운 가상의 유전자 seed의 발현 데이터 값을 생성한다. 여러 개의 seed를 생성하는 경우에는 같은 클러스터에서 같이 나타나는 유전자들의 집합의 크기 순서대로 선택된다.

③ seed를 이용한 군집 클러스터링: ②에서 생성한 seed들을 K-means나 SOM같은 군집 클러스터링의 초기치로 사용하여 클러스터링 한다. 이때 클러스터의 개수는 seed의 개수와 동일하며 클러스터링 과정 중에서 서로 다른 seed가 속한 클러스터가 합쳐지면 그 바로 전 단계에서 클러스터링을 멈춘다.

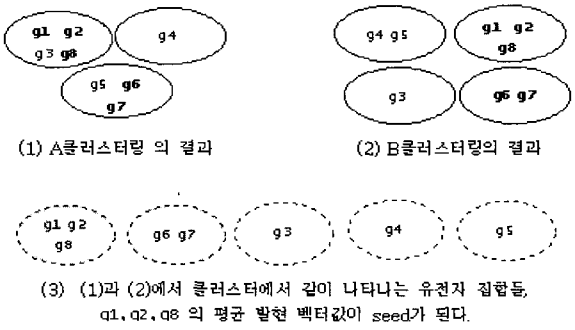

그림 2 Seed 추출 과정

2.1.2 생물학적 온톨로지를 이용한 클러스터의 해석과 평가

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.

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 수치화함으로써 비교 가능하다는 장점이 있으나, 각 클러스터에 속한 유전자 데이터의 생물학적 의미를 반영하지 못한다. 만일 클러스터의 수학적인 평가척도는 나쁘지만, 같은 기능을 하는 유전자들로 묶여있다면 분석의 관점에 따라 잘 묶여진 클러스터로 해석할 수 있기 때문이다.

이에 대한 방안으로 체계화 되어있는 생물학적 온톨로지를 이용하여서 클러스터의 특징을 해석하는 방법이 있다[4]. *GEDA-C*에서는 대표적인 생물학적 온톨로지

인 Gene Ontology(이하 GO)를 이용하여 클러스터의 대표 특징을 파악하고, 대표 특징의 p-value를 계산하여봄으로써 특징의 유의미성을 통계학적인 관점에서 평가하였다.

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 생물학적 온톨로지이다. 여기에 사용된 GO 용어 간에는 부모-자식의 관계가 설정되어 있으며, DAG라는 전체 구조로 되어있다.

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.

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

G : 주어진 종 내에서 전체 유전자의 개수
C : 주어진 GO 용어를 주석정보로 가지는 유전자 개수
n : 클러스터 내 유전자의 개수
k : 클러스터 내에서 주어진 GO 용어를 주석정보로 가지는 유전자의 개수

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 가지지 어렵다는 뜻이다. 즉, 클러스터를 대표하는 GO 용어의 p-value값이 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-value값이 작을수록 클러스터가 잘 묶인 것으로 해석할 수 있다

2.1.3 실험결과

실험 데이터로는 2000년 Mol. Bio에 발표된 "Genomic expression programs in the response of yeast cell to environment changes"에서 공개한 데이터를 사용하였다[5]. 조건 173개에 총 유전자 6152개 유전자 발현 데이터 중에서 missing value가 전혀 없는 유전자 755개를 선별하여 사용하였다. 또한 논문[5]에 따르면 분석결과 비슷한 특징으로 묶이는 17개의 클러스터가 있다. 이를 바탕으로 군집형 클러스터링 방법인 *K*-means와 *GEDA-C*에서 제안한 방법인 seed *K*-means 클러스터링 방법을 비교하였다. 클러스터링에서 유전자 사이의 거리는 모두 유클리디언 거리를 사용하였다. *K*-means의 경우에는 클러스터의 개수는 17개 반복횟수는 100회를 주었다. Seed *K*-means의 경우에는 먼저 complete linkage를 사용하여 Hierarchical 클러스터링 한 후 34개의 하위 클러스터를 생성하였고, *K*-means로부터 34개의 클러스터 생성하였다. 그리고 두 결과로부터 17개의 seed를 생성하고 이를 다시 클러

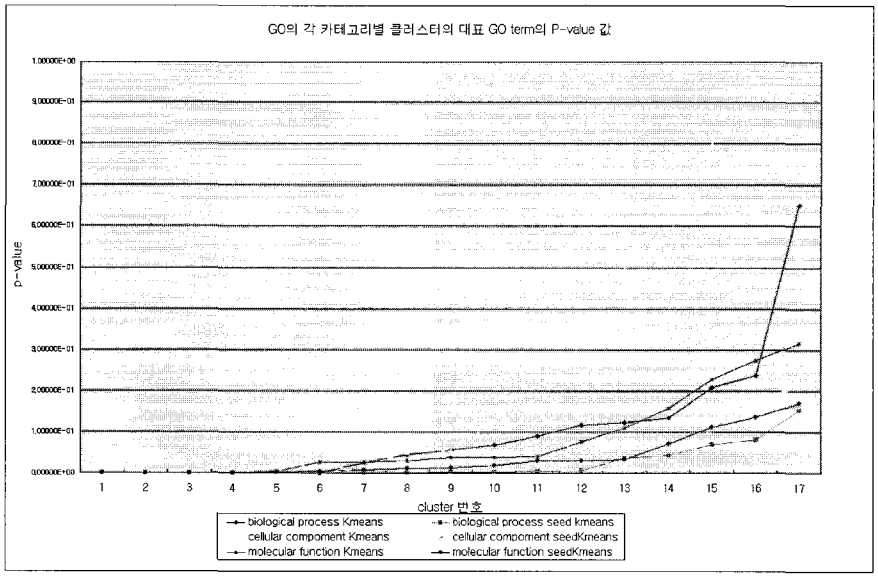

그림 3 GO를 이용한 클러스터링 방법의 비교

스터링 하였다. *K*-means와 Seed *K*-means의 클러스터링 결과는 GO를 이용하여 클러스터별로 대표 GO 용어에 대한 *p*-value를 계산해주는 프로그램[4]을 이용, 이를 비교하여 보았다. 그림 3에서 보이듯이 seed *K*-means 클러스터링에서 생성한 클러스터의 *p*-value값이 GO의 세 가지 카테고리에서 모두 *K*-means 보다 작다. 이는 GO측면에서 클러스터를 해석하였을 때 seed *K*-means 클러스터링이 더 잘 된 클러스터링이며, 더 구체적인 GO 용어로 묶인 클러스터 들을 생성한다고 해석할 수 있다.

대량의 유전자 발현 데이터를 분석할 때 먼저 데이터를 클러스터링 함으로써 유전자 발현 데이터에서 나타나는 특징적인 클러스터들을 파악할 수 있다. 여러 클러스터링 방법 중에서 *GEDA-C*에서 제시한 seed 클러스터링 방법은 클러스터 알고리즘이나 파라메타에 민감하지 않은 유전자 집합들을 클러스터링에 이용하여 클러스터링함으로써 구체적인 특징들로 묶여지는 클러스터들을 생성할 수 있다. 또한 생물학적 온톨로지인 GO를 사용하여 클러스터의 대표 특징을 파악함으로써 기존의 생물학적 기반이 배제된 수치적 클러스터 평가가 아닌, 생물학적 지식에 기반을 둔 클러스터를 해석이 가능하였다.

2.2 단백질정보 분석 기술

2.2.1 단백질 상호작용 네트워크

현재 대부분의 생명 과학 연구는 "하나의 유전자가 하나의 단백질을 만들고 또한 하나의 단백질이 하나의 기능을 수행한다."는 기존의 일차원적인 범위를 벗어나 복잡한 생물학적 기능을 단백질들 사이의 상호작용을 통해 규명하려는데 초점을 맞추고 있다. DNA에 포함된 유전자 정보가 발현되어 최종적으로 생성되는 물질로서 단백질은 다른 단백질과의 상호 유기적인 작용을 통해 신호전달(signal transduction), 세포 생명 주기(cell life cycle), 세포 분화(cell development), DNA 복제(replication), 물질대사(metabolism) 등과 같은 세포의 생리활성 반응을 조절하게 된다. 이 상호작용을 단백질들 사이의 관계로 나타내면 네트워크 형태로 표현된다.

일반적으로 단백질 상호작용 데이터들은 Yeast Two-Hybrid라는 생물학적 실험 방법을 통해 밝혀지고 있다. 또한, 현재 방대한 단백질 상호작용들을 단순한 실험으로 알아낼 수 있는 많은 방법들이 개발됨에 따라 계속해서 축적되는 데이터들을 데이터베이스에서 체계적으로 관리하고 있다.

대표적인 데이터베이스로 PIM(Protein Interaction Map database), BIND(Biological Interaction Network Database), DIP(Database of Interacting Protein), GRID(General Repository for Interaction Datasets) 등이 있다. 이 데이터들은 여러 데이터 마이닝 처리 과정을 통해 새로운 정보가 부여되며 생물학자에게 시각화된다. 이 시각화된 정보를 통해 생물학자들은 새로운 정보를 밝혀내거나 기존에 데이터에 존재하는 오류 등을 제거할 수 있을 것이다. 이 상호작용 데이터는 신약 개발과 같은 고부가가치 산업에 핵심적으로 이용되기 있기 때문에 이들과 관련 시스템 역시 향후 많은 수요가 발생할 것이다.

이를 지원하기 위해 현재 Biolayout, PIMRider, PIVOT(Protein Interactions VisualizatiOn Tool), InterView, Osprey, Cytoscape 등과 같이 네트워크를 자동으로 시각화해 주는 많은 도구들이 개발되어 있다. 이들은 대부분 시각화를 위해 FDP(Force-Directed Placement)를 이용하고 있다. 그러나, 이러한 알고리즘은 방대하고 서로 연결되지 않은 부분 네트워크를 가지고 있는 단백질 상호작용 네트워크를 시각화하기에는 적합하지 때문에 새로운 시각화 알고리즘이 요구되고 있다.

일반적으로 단백질 상호작용 네트워크는 생물체의 종(organism)에 따라 참여하는 단백질뿐만 아니라 이들 사이의 관계가 다를 수 있다. 같은 종에 대해서도 세포 조직(tissue)이나 세포 주기(life cycle)에 따라 여러 종류의 네트워크가 존재할 수 있다. 또한, 하나의 네트워크는 매우 방대한 단백질들과 그들 사이의 복잡한 관계들로 구성되어 있다.

따라서, 단백질 상호작용 네트워크를 위한 시스템은 1) 방대한 네트워크 데이터베이스에서 사용자가 원하는 특정 네트워크를 개념적으로 검색할 수 있어야 하고, 2) 복잡하게 표현된 네트워크를 용이하게 분석할 수 있도록 최적화된 형태로 시각화할 수 있어야 하며, 3) 시각화된 방대한 네트워크에서 사용자가 관심이 있는 특정한 부분 네트워크를 선별적으로 필터링할 수 있어야 한다.

이 시스템은 세포에 존재하는 단백질들 사이의 복잡한 관계들로 표현되는 네트워크들을 사용자가 점진적으로 항해할 수 있도록 지원해야 한다. 이 항해를 위해서는 사용자가 방대한 데이터베이스에서 자신이 원하는 네트워크만을 검색하여 시각화할 수 있으며, 시각화된 네트워크에서 사용자가 관심이 있는 일부 부분 네트워크나 단백질의 구체적인 정보를 점진적으로 탐색해 나갈 수 있어야 한다. 다음은 단백질 상호작용 네트워크 시스템의 요구사항을 나타내고 있다.

□ 단백질 상호작용 네트워크의 검색

기존 네트워크 검색 방법은 단지 종(organism)이나

단백질의 포함여부에 따라 네트워크를 검색하였다. 그러나, 특정 네트워크에 대해 "세포 구성요소(cellular component)", "생물학적 역할(biological process)" 그리고 "분자 기능(molecular function)" 각각에 대한 개념적인 질의를 통해 네트워크를 검색할 수 있는 방법이 더욱 요구되고 있다. 이때, "세포 구성 요소", "생물학적 역할" 그리고 "분자 기능"에 대한 질의는 GO(Gene Ontology)의 용어들로 표현될 수 있으며, 이 용어들은 개념기반으로 네트워크를 검색할 수 있다. 예를 들어, 생물학적 역할에 대한 질의 "apoptosis regulator"에 대해서 개념적으로 "apoptosis regulator"에 속하는 "apoptosis activator", "apoptosis inhibitor" 등과 같은 생물학적 역할을 포함하는 네트워크들도 검색할 수 있다.

1) 단백질 상호작용 네트워크의 시각화

기존 네트워크 시각화는 대부분 FDP 알고리즘을 이용하고 있다. 그러나, 단백질 상호작용 네트워크는 방대한 노드와 에지 또한 여러 개의 부분 네트워크로 구성되어 있다. 따라서, 이 네트워크 특징을 잘 반영할 수 있는 향상된 FDP 알고리즘이 요구되고 있다. 대표적으로 다중 레벨의 FDP 방법(MFDP: Multilevel algorithm for Force-Directed Placement)을 이용한다면, 방대한 단백질들과 복잡한 관계들로 구성된 네트워크를 최적화된 형태로 자동으로 시각화할 수 있다. 또한, 하나의 상호작용 네트워크에 포함된 연결되지 않은 많은 부분 네트워크들에 대해 개별적으로 MFDP를 수행하고 적절하게 재배치할 수 있도록 성능을 향상시켰다.

□ 부분 네트워크 필터링

일반적으로 사용자는 네트워크 전체가 아니라 관심이 있는 특정 부분을 집중적으로 분석하려는 성향을 가진다. 따라서, 먼저 사용자가 방대한 단백질들 중에서 관심이 있는 단백질을 GO(Gene Ontology) 용어를 이용하여 개념기반으로 검색한다. 다음으로, 검색된 단백질들과 일정 거리 안에 연결되어 있는 부분 네트워크를 새로운 윈도우에 시각화하여 사용자가 관심 있는 부분을 집중적으로 분석할 수 있도록 지원해야 한다.

2) 경로 탐색 및 네트워크 단백질 정보 참조

사용자는 두 단백질 사이의 경로들에 포함된 단백질들에 관심이 있을 수 있다. 따라서, 두 단백질 사이의 최단 경로를 포함한 모든 경로를 쉽게 탐색할 수 있어야 한다. 또한, 이 경로들 중에 특정 단백질을 경유하는 경로들만을 탐색 할 수도 있어야 한다. 이 경로에 있는 단백질에 대한 자세한 정보는 Swiss-Prot과 같은 다른 데이터베이스와 연결된 링크를 통해 참조할 수 있어야 한다.

이 단백질 상호작용 네트워크는 생물학적인 관점에서 단백질의 기능을 예측하기 위한 중요한 정보로 이용된다. 즉, 하나의 단백질의 기능은 이 단백질과 상호작용을 하는 단백질의 기능과 유사하다는 기본 가정에 따라 기존에 기능이 잘 알려진 단백질에 대한 정보를 통해 이 단백질과 상호작용을 하는 여러 미지의 단백질들에 대한 기능을 예측할 수 있을 것이다. 또한, 단백질 상호작용에 대한 이해는 신약개발에 있어 매우 중요하게 인식되고 있다. 즉 질병의 원인이 단백질들 사이의 상호작용 네트워크에서 발생하게 됨으로, 상호작용 네트워크를 이해하고 이를 조절한다면 획기적인 신약을 개발할 수 있을 것이다.

2.2.2 단백질 상호작용 관계를 이용한 단백질 기능 예측

초기의 바이오인포매틱스 연구는 유전자 서열 정보 등 방대한 양의 새로운 생물학 데이터들을 저장하고 분석하기 위한 데이터베이스 개발에 초점을 맞추었지만, 현재의 많은 연구들은 미지의 단백질에 대한 구체적인 기능을 예측하는데 관심을 가지고 있다. 특히, 단백질 상호작용 데이터로부터 특정 단백질에 대한 신뢰성 있는 기능 예측을 위해 효과적인 프로티오믹스의 계산 모델 개발이 절실하게 요구되고 있다. 이 계산 모델을 이용한 단백질 기능 예측은 매우 부가가치가 높은 신약개발이나 의료진단에 이용됨으로 생물학적 실험에 요구되는 비용을 현격하게 감소시킬 수 있다.

현재, 국내 대학 및 연구소 등에서 진행 중인 단백질 기능 예측 관련 연구는 아직 초기단계에 불과하며, 실제 국내에서 특정 유기체에 대한 대규모 단백질 기능 예측 모델에 관한 연구 및 개발된 사례는 알려지지 않고 있다. 또한, 실험실에서 병의 진단, 제약 관련한 한정된 수십 개의 단백질의 기능 예측이 실험으로부터 행하여지고 있으며, 대규모 단백질 관련 데이터를 이용한 예측 모델에 대한 연구 및 개발은 없는 실정이다.

해외에서는 단백질 기능 예측의 모델링에서 단백질을 발현한 유전인자의 서열(sequence), 표현형 및 구조(structure)의 상동관계(homology) 데이터를 이용한 방법 및 단백질-단백질 상호작용 데이터를 이용하는 방법 등이 계산이론의 응용에서 연구되었으나 실험 데이터에 의하면, 상동관계를 이용하는 방법보다는 단백질들 사이의 상호작용 여부를 이용한 방법이 높은 신뢰(reliability)을 갖는다고 알려져 있다.

단백질-단백질 상호작용 데이터를 이용하는 방법 중 그래프 기반 계산 이론(graph-based computational theory)의 응용은 단백질의 기능 예측 문제에서 많은 기능이 알려지지 않은 단백질의 기능 예측에 적합한 모

델이다. 현재까지 진행된 기술 수준을 고려할 때 단백질 기능 예측에 대한 연구개발은 효과적인 알고리즘의 응용 단계에 있으며, 밝혀진 특정 유기체에 대한 단백질 및 그 외 관련 생물학적 사실로부터 수집된 데이터를 대상으로 시뮬레이션(simulation)을 이용한 기능 예측 모델의 응용에 적합성 검증이 진행되고 있다.

단백질 기능 예측 도구는 단백질-단백질 상호작용 데이터 분석 시 주요한 기능을 제공하며, 예측 결과를 이용하여 다양한 세포 내 생물 현상을 시뮬레이션 할 수 있는 환경을 제공하게 된다. 또한 이러한 기술은 비교생물학(comparative biology)및 유전체학(genomics)을 처리하는데 있어 핵심이 되는 기술이다. 상호작용 데이터를 이용하여 미지(unknown function)의 단백질 기능 예측을 하는 방법은 현재 우수한 외국 바이오 업체에서 중요한 연구 분야로 인식되고 있다. 다음은 상호작용 관계를 이용한 단백질 기능 예측에 이용되는 대표적인 모델이다.

- □ Neighbor-Counting 예측 모델: Guilt-by-Association 개념 기반 예측 모델로 그래프 계산 이론을 이용한 인스턴트 기반 알고리즘(instance-based algorithm)에 속하는 예측
- □ x2-통계치 예측 모델: 하나의 단백질은 여러 개의 기능 및 기능별 바인딩 파트너를 가질 수 있다는 사실을 모델에 응용한 것으로, 단백질 상호작용 네트워크로부터 위상(topology) 정보를 이용한 정해진 깊이(depth)내에서 기능별 x2-통계치를 이용하여 기능 예측
- □ Interaction Generality 예측 모델: Yeast-two Hybrid 방법으로 얻어진 단백질 상호작용의 약 50%정도는 False-Positive로 평가되고 있기에, 단백질 상호작용의 신뢰성을 평가하기 위해 두 단백질간의 Correlation Expression과 Interaction Generality를 기능 예측

바이오인포매틱스의 가장 큰 목적 중에 하나는 여러 생물학적 데이터를 기반으로 단백질의 기능을 예측하는 것이다. 따라서, 단백질 상호작용 관계 데이터를 이용한 미지의 단백질 기능 예측은 매우 중요한 연구 분야로 인식되고 있다. 또한, 신약 개발에서 목표 단백질(target protein)을 예측할 때 매우 유용하게 이용될 수 있다.

2.2.3 단백질 가시화 도구 및 분류 기술

단백질 연구에 있어서 단백질 구조는 매우 유용한 정보를 제공한다. 왜냐하면, 단백질의 기능은 단백질의 구조와 밀접한 관련이 있기 때문이다. 이러한 구조정보를 잘 축적하고 있는 데이터베이스로 RCSB(Research Collaboratory for Structural Bioinformatics)의 PDB가 있다.

Protein Data Bank (PDB; http://www.rcsb.org/pdb/)는 단백질, RNA와 같은 생물학적 고분자들의 구조데이터(structure data)를 집적해놓은 일종의 도서관과 같은 곳이다. 1971년, Brookhaven National Laboratories(BNL) 에서 고분자의 crystal structural data를 모으기 위해 만들어졌으며, 1980년대에 핵자기공명법(Nuclear Magnetic Resonance;NMR), X선 회절법(X-crystallography) 등의 분자 구조 결정 실험 방법의 발전으로 인해 축적된 고분자 구조 data의 양이 늘어났다. 이와 같이 날로 증가하는 단백질 구조를 가시화하고 조작하기 위해서는 분자(단백질도 고분자의 일종) 구조 가시화 도구가 필요하다. 또한, 이들 구조 정보를 이용하여 단백질의 기능을 분석하기 위한 방법으로 분류기술이 요구된다.

1) 단백질 모델링 및 가시화 도구(Protein Modeling & Visualization Software)

분자의 위치 좌표 정보(coordinate data)를 사용하여 모니터 상에서 단백질이나 RNA와 같은 고분자의 형태를 나타내고, 더 나아가 그것의 조작을 가능하게 하는 프로그램이다. 이때 필요한 위치 좌표 정보는 일반적으로 분자학적 실험(NMR, X-Crystallography)의 결과로 얻을 수 있다.

이러한 가시화(Modeler) 프로그램들은 일반적으로 분자의 형태를 사용한 분자 반응의 시뮬레이션(simulation)(분자의 기능을 알아보기 위한), 논문 게재를 위한 분자 그림을 얻고자 할 때 사용된다. 또 몇 가지 인위적인 조작을 통해 단백질 등의 고분자가 생체 내에서 변형이 생겼을 때 그 기능변화를 예측해볼 수도 있다.

과거에는 이러한 작업을 위해 몇 대의 수퍼컴퓨터가 필요했으나 오늘날에는 CPU의 기능 향상과 가격의 인하로 인해 펜티엄(Pentium), 파워-맥(Power-Mac) 환경에서도 이러한 프로그램의 구현이 가능하며, 많은 프로그램들이 개발되고, 인터넷 등을 통해 공개되고 있다.

본 연구에서는 단백질의 구조를 다양한 그래픽스 기법들을 적용하여 실감 있게 표현하는 기술들을 개발하였다. 이를 위하여 범용성의 OpenGL기반과 Windows 계열 PC에 최적화된 DirectX 기반의 가시화 도구를 개발하였다. 본 연구개발을 다양한 단백질 모델에 의한 가시화가 가능하고, 2차 구조 가시화 기능을 강화하였으며, 단백질 표면 가시화 기능과, 단백질 표면 및 내부 가시화 기능, Trace 기능(단백질의 N 말단(N-Terminal)에서 C 말단(C-terminal)으로 구조를 동적으로 가시화하는 기능) 등의 고급 수준의 가시화 기능을 제공한

다. 또한, 단백질 분류 프로그램과 연계하여 단백질 구조를 비교하여 가시화 할 수 있는 기능을 제공한다. 이는 단백질의 구조의 유사정도를 가시화 도구를 사용하여 확인할 수 있는 강력한 기능이라 할 수 있다. 이 가시화 도구는 앞으로 본 연구 과제를 통하여 수행될 단백질 구조 관련 분석기능들과 연계되어 더욱 세밀하고 자세한 가시화를 제공할 것이다.

2) 단백질 분류 기술(Protein Classification Technique)

단백질 연구에 있어서 단백질 구조에 의한 분류를 통하여 구조와 기능의 상관관계를 규명하려는 노력들이 있어왔다. 이러한 노력들은 단백질 구조 분류연구로 이어졌고 이러한 분류의 결과, 단백질 구조 분류 데이터베이스들이 생겨났다. SCOP(http://scop.berkeley.edu/)과 CATH(http://www.biochem.ucl.ac.uk/bsm/cath/) 등이 그것이다.

단백질 분류 시 기존의 방법은 분류기준을 미리 정하여 고정시킨 후 단백질을 분류하기 때문에 생물학자나 사용자가 단백질을 새로운 형태로 분류하기 불가능하다. 이러한 단점을 보완하고 단백질 구조로부터 새로운 정보들을 찾아내기 위하여 본 연구에서는 사용자가 재구성 가능한 동적 분류 기술을 제안하였다. 이를 이용하면, 생물학자나 사용자가 분류할 대상 단백질을 선택하고, 분류할 기준을 특정한 목적에 맞게 선택하여 조합하는 형태로 단백질을 분류할 수 있다. 단백질 구조 유사성을 판단하기 위하여 객관적인 유사도 측정방법을 조합하여 사용자의 목적에 맞는 단백질 분류기를 제작할 수 있다.

현재까지 분류에 적용된 기준들로 단백질 서열, 단백질 2차 구조 개수, 단백질 2차 구조 서열, 라마찬드란 맵, 3D 백본 결합 분포, 회귀분석(Regression), 단백질 2차구조의 구조적 특징 등이 있다.

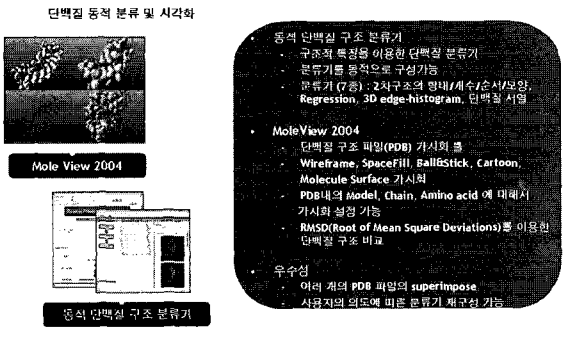

그림 4 개발된 단백질 구조 가시화 도구와 동적 단백질 구조 분류기

지금까지 연구내용은 단백질 전체구조에 대한 특징들을 이용하여 분류를 수행하였다면, 앞으로의 연구계획은 단백질이 특정 단백질 혹은 신약 후보 물질등과 작용하는 부위에 대한 구조적 분류에 대하여 연구를 진행하고자 한다. 이와 병행하여, 가시화도구와 연계하여 활성부위에 대한 상세한 분석이 가능하도록 도구를 개발할 예정이다.

이러한 기술이 현장에 적용되면 기존의 고정된 분류방법과는 달리, 다양한 분류기의 조합을 통하여, 단백질을 분류하였기 때문에, 생물학자의 실험과 요구사항을 빠르게 충족시킬 수 있을 것으로 기대한다. 또한, 특정한 단백질 그룹을 다양한 방법에 따라 분류함으로써, 분류방법에 대한 단백질의 새로운 특성 확인할 수 있다. 끝으로, 구조에 기반한 단백질 연구는 기존의 단백질 서열에 관한 연구보다 단백질에 관한 더 많은 생물학적 정보를 가지고 있기 때문에, 신약개발 등의 연구 분야에 있어 더욱 중요하고, 핵심적인 연구 분야가 될 것으로 예상된다.

2.3 자연어처리 접목 기술

바이오산업의 고속 발달로 인해 그와 관련된 다양한 형태의 바이오 데이터가 대량으로 산출되고 있다. 바이오 문헌 데이터에는 연구에 대한 고수준의 분석 결과가 기술되어 있고 이러한 문헌들이 지속적으로 산출되고 있는 상황이기 때문에 이를 대상으로 한 자동적인 정보추출은 필수 불가결한 요구사항으로 부각되고 있다. 그러나 현 시점에서 바이오 문헌을 대상으로 한 정보추출은 기초단계의 연구가 진행되고 있는 수준이다. 바이오 문헌을 대상으로 한 정보추출이 실현될 경우 추출된 정보 자체만으로도 하나의 의미 있는 지식자원을 구성할 수 있을 뿐만 아니라 이를 각종 기초 자료와 결합함으로써 기초 자료의 정보 가치를 높일 수 있게 될 것이다.

2.3.1 연구 이슈

바이오 텍스트 마이닝에서 중점적으로 다루어지고 있는 연구 이슈들로는 생물학적 개체명 인식, 생물학적 개체 간의 관계 인식 및 네트워트 구성 등이 있다. 생물학적 개체명 인식은 생물학적 문헌들에서 유기체명, 단백질명, 유전자명 등과 같은 생물학적 개체들을 지칭하는 이름들을 추출하는 것으로, 이는 생물학 문헌의 정보 주체를 파악하기 위한 목적으로 이루어지는 바이오 텍스트 마이닝의 가장 기본이 되는 정보 추출 단계이다. 생물학적 개체 간의 관계 인식은 개체명 인식의 결과를 기반으로 이들 간에 존재하는 생물학적으로 중요한 의미를 갖는 관계들을 추출하는 단계이다. 이는 개체명 인식 단계에서 얻은 문헌의 정보 주체들을 의미적으로 연결함으로써 특정 정보와 관련된 정보들을 보다 손쉽게 분석할 수 있도록 하기 위한 목적으로 수행된다. 생물학적 개체 간에는 단계적인 상호작용이 일어나는 것이 일반적이기 때문에 문헌으로부터 이러한 단계적인 흐름들을 분석해내는 것도 중

요하다.

2.3.2 연구 내용

1) 생물학적 개체명 인식

생물학적 개체명 인식 방법은 크게 규칙 기반 방식과 통계 기반 방식으로 나누어 볼 수 있다. 규칙 기반 개체명 인식의 경우는 생물학 분야에서 사용되는 개체명들을 사전 형태로 구성하고 생물학 분야에 적합한 개체명 인식 규칙들을 정의한 후에 생물학 문헌들을 대상으로 사전 및 개체명 인식 규칙을 패턴 매칭 형태로 적용시켜 개체명을 인식해내는 방식이다. 이 방식의 경우는 새롭게 출현하는 개체명들이나 기 정의된 규칙들을 만족하지 못하는 변형된 형태의 다양한 개체명들을 인식하지 못한다는 단점을 갖는다. 통계기반 개체명 인식은 우선적으로 학습 코퍼스를 구축하고 이를 이용해 다양한 통계적 학습 알고리즘들을 적용해 개체명을 인식하는 방식으로 이 방법의 경우는 대용량 학습 코퍼스를 구축하는데 많은 비용과 시간이 소모된다는 단점이 있다. 이를 위하여 ETRI 바이오정보연구팀에서는 다양한 생물학 도메인의 태깅 코퍼스를 구축중이다. 아울러 규칙기반 및 통계기반의 생물학적 개체명 인식기를 개발하고 이들의 성능 개선을 위한 연구를 수행중에 있다.

2) 생물학적 관계 인식

생물학적 개체 간의 관계 인식은 개체명 인식의 결과를 기반으로 이들 간에 존재하는 생물학적으로 중요한 의미를 갖는 관계들을 추출하는 2차적인 정보추출 단계이다. 이는 개체명 인식 단계에서 얻은 문헌의 정보 주체들을 의미적으로 연결함으로써 특정 정보와 관련된 정보들을 보다 손쉽게 분석할 수 있도록 하기 위한 목적으로 수행된다. 예를 들면, 단백질이 서로 어떻게 연관되는지를 아는 것은 분자경로의 개발에 중요한 정보로 이용된다. 생물학적 개체들 간의 관계는 개체들의 기능적 분류에 따라 다양한 관계들이 있을 수 있기 때문에 일반적인 텍스트 마이닝에서 사용되는 텍스트 단위간의 단순 공기 빈도(co-occurrence frequency)를 이용한 관계 인식 기법으로는 이러한 다양한 관계를 제대로 분별해낼 수 없다. 따라서, 생물학적 개체들 간의 기능적인 관계를 나타내는 어휘 집합이나 생물학적인 기능들을 계층적으로 구조화한 온톨로지 등을 활용한 방법들이 주로 이용되고 있다. 생물학적 개체들 간의 관계를 인식하기 위한 방법은 크게 규칙 기반 방식과 통계 기반 방식으로 분류할 수 있다.

규칙 기반 접근방법은 주로 자연어처리 기법을 이용하여 개체 간의 관계를 인식한다. 먼저, 사전 및 규칙을 이용하여 생물학적 개체명을 인식하고 태깅 또는 단순 패턴을 이용하여 문장의 구조를 분석한다. 그 결과 얻어진 정보와 전문가가 수동으로 미리 정의한 서술 동사 위주의 패턴을 이용하여 생물학적 관계를 인식한다. 이에 반해, 통계 기반 접근방법은 주로 공기 빈도를 이용하여 관계를 인식한다. 일반적으로 생물학적 개체명을 인식한 후, 함께 출현하는 생물학적 개체명들과 이들 간의 관계를 인식한다. 이렇게 인식한 결과, 특정 임계값 이상의 관계성을 갖는 생물학적 관계를 필터링하여 최종적으로 개체명과 그들 간의 관계를 결정한다.

2.3.2 연구 활용

인식된 개체명은 관계인식을 통하여 개체간의 관계로 표현 될 수 있다. 이 데이터는 바이오 네트워크의 구축 및 기 구축된 데이터의 검증에 활용 가능하다. 바이오 문헌은 유전자 및 단백질의 위치와 구조 정보, 중요 유전자 패턴 정보 등의 전반적인 생물정보학 연구 분야에 대한 기반 지식을 포함하고 있어 생물학적 정보 및 지식 획득을 위한 주요한 자원으로 활용될 수 있다. 중요한 지식으로는 생물학적 중요 개체명과 이들의 관계를 비롯하여 이들 개체간의 조절 기작 및 누가 무엇을 연구하는가에 대한 정보 등이 있을 수 있다. 이러한 바이오 문헌을 대상으로 한 바이오 텍스트 마이닝 기술은 생물학적 사전 및 축적된 다양한 바이오 정보들을 개념적으로 연결시켜 새로운 정보를 제공하는 바이오 지식 생성도구의 핵심 기술로 쓰여 질 수 있다. 향후 개체명과 관계 인식에 관한 연구에서는 새로운 도메인으로의 확장성을 위해 사전 및 명명법에 대한 의존도를 낮추고 자동으로 개체명과 관계를 인식하는 기법이 요구된다. 또한 텍스트 마이닝 도구의 성능을 평가할 수 있는 공인된 평가집합의 구축이 시급하다. 바이오 문헌으로부터 얻은 데이터를 대규모의 실험 데이터 분석 결과와 결합하여 보다 정확한 정보를 제공하는 방법에 관한 연구도 이루어져야 할 것이다.

2.4 웹서비스 기반 생물정보 접목 기술

생명과학이 사회와 인간의 삶에 미치는 영향력은 제약·의료 등과 같은 직접적인 연관 분야뿐만 아니라 다른 산업분야에 파생되는 효과가 매우 크다. 그러한 산업분야중의 하나인 정보기술(IT, Information Technology) 분야는 생명과학의 진보에 따라 생성된 많은 량의 정보들을 효율적으로 제어하고 관리하기 위한 요구로 자연스럽게 대두되었으며, 이러한 일련의 정보기술들을 생물정보학(Bioinformatics)이라고 한다.

생물정보학은 인간 유전체(Genome) 프로젝트와 더불어 발달한 분야로 생물학적 연구에서 나오는 각종 데이터의 분석 및 관리를 지원하여 새로운 생물학 데이터를 창출하거나 기존 연구에 도움이 되는 관리의 개념을 포함한다. 이때, 필수적으로 활용되는 생물정보 데이터

베이스는 다양한 형태로 여러 서버에 산재되어 있는데, 생물정보 분석을 위해서는 여러 데이터베이스에 대한 복합적인 접근과 활용을 필요로 한다. 그러나 실제 생물정보 데이터는 데이터 공급자들 간의 상이한 데이터베이스 구조, 포맷 형식, 데이터 정의 용어의 상이함, 구현 언어의 차이 등으로 인해 통합에 많은 어려움을 안고 있다.

본 절에서는 이러한 어려움을 완화하고 실제 데이터의 활용도를 높이기 위해 필요한 생물정보 통합 시스템의 기반이 되는 생물정보 수집 시스템에 관한 것으로, 웹서비스 개념을 도입하여 이종 생물 정보 데이터베이스에 존재하는 다양한 데이터를 수집하는 시스템에 대한 것이다. 특히, 여러 형태의 생물정보 데이터베이스에 대해 생물학자가 원하는 형태로 결과를 취합할 수 있도록 생물정보 데이터를 XML화하기 위한 명세 언어 개발 내용을 중점적으로 소개하고자 한다. 이를 기반으로, 다양성과 유동성을 갖는 생물 정보의 특성을 고려하여 웹서비스 기반으로 통합함으로써 보다 능동적으로 데이터에 접근할 수 있는 방법을 함께 소개하고자 한다. 시범 대상이 된 생물정보 데이터베이스는 NCBI의 GenBank, Gene DB와 AmiGO, UniProt이다.

2.4.1 관련연구

1) 웹서비스 기반 생물 정보 통합

웹서비스란 인터넷 환경에서 다양한 OS, 미들웨어, 애플리케이션 구현 방법 등에 상관없이 표준으로 정의된 언어(XML)와 통신/표현 방법(SOAP, WSDL, UDDI)을 사용하여 서비스 단위의 통합을 이루어 유연하고 능동적인 웹 애플리케이션 수행환경을 제공하는 기술을 말한다.

웹서비스는 실행 플랫폼과 독립적인 공통된 인터페이스와 접근/통신 방법을 제시하고 있으므로, 산재된 생물정보 데이터를 통합하기 위해 적절하게 활용될 수 있다. 이 때문에 Bio-MOBY[2], DDBJ[4], MyGrid Project [5] 등과 같은 웹서비스 기반 생물정보 통합에 관한 연구가 진행 중이다. 이외에도, 통합 환경을 제공하지는 않지만, 개별 데이터베이스를 웹 서비스화 하는 경우도 있다. XEMBL[6]은 EBI의 EMBL 염기서열 데이터베이스를 검색, 조회하는 서비스를 제공한다. 즉, SOAP에 ID와 포맷에 관한 인자 값을 받아 유전자 염기 서열을 XML 문서로 반환한다.

한편, 생물 정보 데이터 기술에 대한 표준화 작업의 일환으로 BSML[3]을 정의하여 데이터 기술에 활용하려는 연구도 진행되고 있다.

2) 생물 정보 수집

생물정보의 수집은 기존의 웹 정보 추출 기능의 확장된 형태이다. 일반적으로 정보 추출 시스템은 html 문서로부터 특정 정보를 패키징 하여 XML 형태의 구조화된 데이터로 변환하는 래퍼(wrapper)를 사용한다. 래퍼는 데이터 소스 자체의 형태적 변경이 있을 시 사용자가 매번 데이터 소스에 대한 추출 프로그램을 새로 작성하지 않고 간단한 규칙 등의 기술만으로 자동으로 데이터 소스에서 정형화된 데이터를 추출하는 기능을 하는 장점이 있다.

XWRAP[1][8]은 대표적인 래퍼중 하나로 html 페이지의 각 태그를 오브젝트화 하여 추출 대상을 규정하고, 제공되는 인터페이스로부터 사용자의 선택에 의해 xml 문서를 자동으로 생성한다. XWRAP Composer [7]는 XWRAP을 생물 정보 도메인에 적용하기 위한 시스템으로 프로모터 모델을 생성하고 확장하기 위해 구축되었다. 시스템의 역할은 마이크로어레이를 분석하여 클러스터링 하고 새로운 후보 유전자에 연결할 공통 프로모터를 검색하는 두 단계로 나뉘어 수행된다. 즉, 마이크로어레이 데이터의 수집을 위해 CLUSFAVOR에 대한 래퍼를 생성하고, 프로모터의 검색을 위해 NCBI와 BLAST에 대한 추가의 래퍼를 생성한다. 특히, 래퍼 생성을 위해 자체의 인터페이스 언어를 디자인하였고, 또한 각 소스로부터의 xml 데이터로의 통합을 위해 스크립팅 언어의 활용도 가능하다. 아울러, 제공된 언어에 대한 사용자의 활용성을 높이기 위한 편리한 사용자 인터페이스를 제공하는 것이 특징이다.

2.4.2 설계 및 구현

1) 정보 통합 웹서비스를 위한 구조

본 절에서 설명하는 웹서비스 기반 생물정보 통합 시스템의 개념도는 그림 5와 같다. 그림 5는 생물 정보 데이터를 제공하는 각종 데이터베이스 서비스와 서비스를 관리하는 레지스트리서버(IBM UDDI), 그리고 실제 서비스를 사용하는 생물 정보 분석 클라이언트로 구성된다.

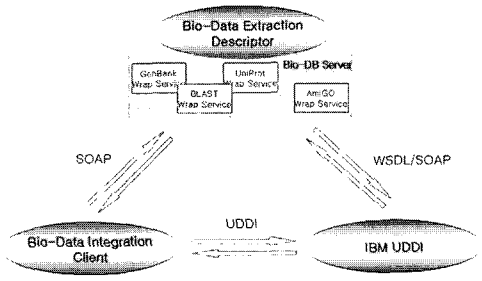

그림 5 생물정보 통합 웹서비스

생물 정보 수집 모듈 생성기(Bio-Data Extraction Descriptor)는 생물정보학적 요구가 있는 각종 데이터

베이스에 대해 데이터 추출 규칙을 명세하고 규칙을 바탕으로 각 데이터베이스에 대한 데이터 수집 서비스를 생성한다. 이들 서비스가 생성되면 WSDL로 각 데이터베이스 수집 서비스의 내용을 작성하여 IBM의 UDDI 레지스트리에 서비스를 등록하고 이후로 서비스 가능한 상태가 된다.

보편적으로 생물학 실험 및 검증의 특성상, 사용자인 생물학자는 하나 이상의 생물정보 데이터베이스에 대한 접근을 요구한다. 따라서, 서비스 사용자는 UDDI 서버에서 사용 가능한 서비스에 대해 원하는 서비스를 검색, 이를 이용하여 데이터를 수집하고 통합하게 된다.

현재 생물정보학 분야의 데이터베이스들은 웹서비스 환경이 갖춰져 있지 않거나 웹서비스의 초보단계에 있으므로, 사용자의 서비스 요구에 적절하게 대응할 수 없다. 따라서, 그림 5의 개념도에서는 생물 정보 수집 모듈 생성기가 각종 데이터베이스에 대한 서비스를 생성하여 실제 웹서비스를 제공하지 않는 데이터베이스에 대해서도 서비스가 가능하도록 하여 일관된 서비스를 이용할 수 있도록 하였다.

2) 서비스 생성을 위한 명세 언어

바이오정보연구팀에서 정의한 래퍼 명세 언어(Wrapper Description Language, 이하 명세 언어)는 데이터베이스에서 추출하고자 하는 각 데이터 항목에 대해 추출 방법과 항목의 구조를 정의할 수 있는 언어이다. 명세 언어는 그림 1의 생물 정보 수집 모듈 생성기에서 각 데이터베이스 서비스를 기술하기 위한 목적으로 사용된다. 명세 언어 문서는 데이터 선언부와 추출 규칙 명세부, 질의 처리를 위한 오퍼레이션부로 구성된다.

데이터 선언부는 XML 데이터 모델의 Element, Attribute에 대응되는 형태로 데이터를 기술할 수 있으며, Element들간의 관계성을 포함한 nested Element 구조도 자유롭게 표현할 수 있다. 이때, 데이터 선언부에 정의된 각 항목들은 고유의 타입을 가진다.

추출 규칙 명세부는 웹상의 생물 정보 추출을 위해 많이 이용되는 정규 표현(Regular Expression)을 사용하여 html 태그를 포함한 문서에서 추출해야 할 데이터에 대한 추출 규칙을 정의한다.

오퍼레이션부는 실제 정보 제공 사이트에 접근하여 추출 대상 웹문서를 가져오기 위해 필요한 행위를 정의한다. 각 사이트는 기본적으로 키를 위한 검색과 키 이외의 검색 방법을 제공하는데 오퍼레이션부에서는 두 가지 검색 방법 모두를 위한 처리 방식을 기술하여야 한다. 실제로 오퍼레이션부에는 접근을 위한 URL과 cgi 파라미터 등에 관한 내용을 포함한다. 한편, 각 부에 사용할 데이터의 할당을 위해 변수를 사용할 수 있다.

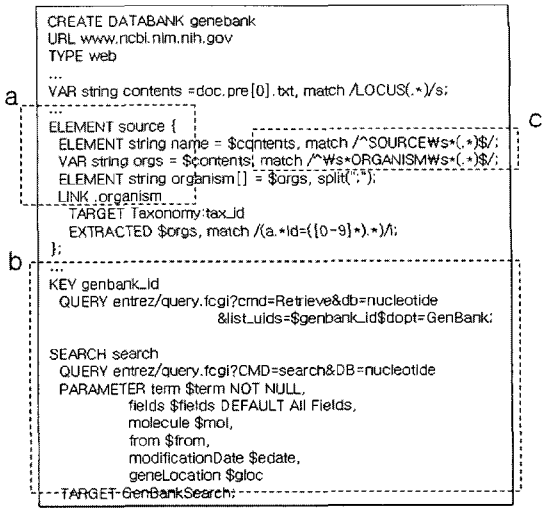

그림 6 NCBI GenBank의 명세언어문서 예(a.데이터선언부 b.추출규칙 명세부 c.오퍼레이션부)

그림 6는 NCBI GenBank 데이터베이스를 명세 언어로 기술한 예이다. 'source'는 하위 element를 포함하는 복합 element이다. 추출 규칙은 'match' 구문 이하에 정규 표현으로 기술되었다. 'LINK'는 html문서에서 'href'와 유사한 역할을 하여 다른 데이터베이스에 대한 참조를 정의한다. KEY와 SEARCH는 두 가지 형태의 검색 방법을 정의한다. KEY의 경우 'genbank_id'를 cgi 파라미터로 하여 URL과 QUERY를 조합하여 해당 사이트에 접근하기 위한 질의를 구성하게 된다. SEARCH의 경우 'field', 'molecule' 처럼 파라미터로 정의된 항목들은 URL + QUERY 구문 이후에 삽입하여 실행할 질의를 작성한다. KEY에 대한 검색 결과는 데이터베이스에서 단지 하나의 항목만이 존재하지만, SEARCH에 대한 검색 결과는 하나 이상의 인스턴스가 검색될 수 있으며 이때 이들을 변환하기 위한 추가의 명세언어 규격이 필요하며 그림 6의 'search'는 TARGET절의 'GenBankSearch'가 추가로 정의되었음을 알 수 있다.

그림 6과 같이 명세 언어로 기술된 각 데이터베이스 서비스들은 사이트에서 제공하는 데이터 항목의 모든 데이터들을 포함한다. 그러나, 실제 사용에 있어 전체 데이터는 네트워크에 부하를 줄 수 있고, 사용자의 측면에서는 원하지 않는 부가적인 데이터까지 받아서 처리해야 하는 단점이 있다. 명세 언어로 실제 수집할 데이터를 한정할 경우 이런 문제를 해결할 수 있으며, 이를 위해 XQuery를 사용한다.

3) 서비스를 위한 통신 메시지

데이터베이스에 대한 명세가 이루어지면 추출 모듈 생성기는 명세 문서를 파싱하여 자동으로 자바코드 형태의 정보 수집 모듈을 생성한다. 이때 정보 수집 모듈은

웹서비스를 위해 필요한 메시지 규약을 포함하며, 서비스 등록을 위한 WSDL 내용도 같이 생성된다.

a. Input

```
<Input>
  <query type="key">NM_417264</query>
  <return type="element">GenBank</return>
</Input>
```

b. Output

```
<?xml version="1.0" encoding="UTF-8"?>
<Output>
  <GenBank>
    <locus>
      <access_num>NM_018850</access_num>
      <length>5623</length>
      ...
    </locus>
    ...
    <source>
      <source_name/>
      <organism link="Taxonomy:9606">Homo sapiens</organism>
    </source>
    <references>
    ...
    </references>
  ...
  </GenBank>
</Output>
```

그림 7 SOAP의 입출력 메시지 예

각 데이터베이스에 대한 추출 모듈은 자동으로 생성되므로 각 수집기 모듈에 접근하기 위한 메시지 규칙은 데이터베이스에 독립적으로 작성된다. 실제 데이터에 접근하기 위해서는 오퍼레이션부에서 질의할 질의어가 필요하며, 사용자가 요구할 수 있는 데이터부의 단위 정보 element와 attribute이름이 요구된다. 입력이 주어지면 데이터베이스를 검색하여 결과를 수집하고 사용자가 지정한 element나 attribute를 XML 문서화하여 SOAP 메시지로 포함 후 반환한다. 각 서비스의 통신에 이용되는 SOAP의 입출력 메시지의 형태는 그림 7과 같다.

질의 형태는 'query' element의 type에 입력되며 질의어는 'query' element의 값이다. 반환 결과는 'return' element의 값이다. 그림 7의 b에서 실제 GenBank element가 반환되는 것을 볼 수 있다.

4) 생물 정보 검색 클라이언트

앞서 정의에 따라 생성한 서비스의 검증을 위해, 생물학 실험에 의해 산출된 마이크로어레이 데이터로부터 여러 데이터베이스 서비스를 거쳐 관련된 유전자에 대한 UniProt 데이터를 획득하는 일련의 단계를 구현하였다. 그림 8은 실험데이터 검증을 위해 사용되는 생물정보 데이터베이스와 이들의 처리 순서도이다.

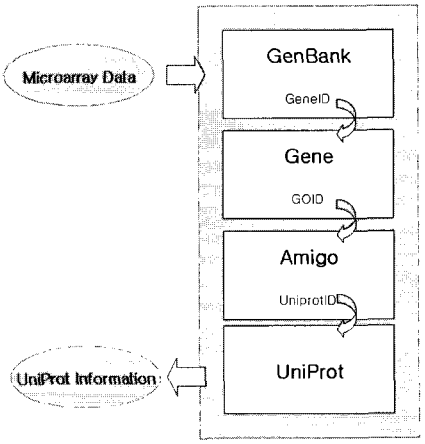

그림 8 클라이언트 처리 순서

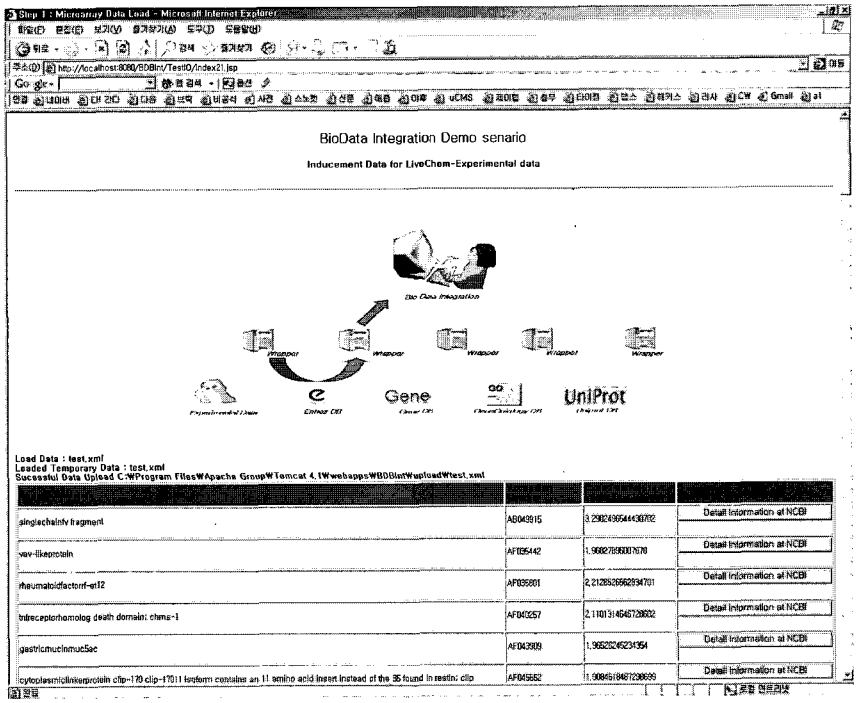

그림 9 웹서비스 클라이언트 예

마이크로어레이를 분석하여 얻어진 유전자에 대한 GenBank의 Accession ID를 GenBank에 질의하여 해당 유전자의 정보를 가져온다. 그리고, 해당 유전자에 참조된 Gene ID를 얻어 Gene DB로부터 Gene 정보를 가져온다. 다음 단계로 해당 유전자와 관련 있는 유전자를 확장하기 위해 Gene Ontology의 정보를 추가로 가져와, 같은 같이 참조되는 다른 유전자들에 대해 이들에 링크된 UniProt의 정보를 최종적으로 사용자에게 반환한다.

생물학자가 접근하는 GenBank, Gene, Amigo, UniProt 서비스들은 명세 문서에 의해 기술되어 생성된 후 IBM의 UDDI에 등록된 상태이다. 클라이언트는 단계별로 서비스에 접근하여 해당 서비스에 request하고 결과를 받아 각 단계를 처리한다.

그림 9는 실제 구현된 웹서비스 예제 사이트이다.

이와 같이 본 절에서는 이종의 생물정보 데이터베이스에 대한 일관성 있는 접근 방법을 제공하기 위해 각종 데이터베이스로부터 데이터를 수집하여 제공하는 데이터 수집 웹서비스 서버와 이들을 이용하여 실제로 생물학 정보를 검증하는 웹서비스기반 통합 예를 구축하였다.

본 절에서 보인 시나리오와 같이, 실험 및 결과 분석을 위해서는 생물정보 데이터의 다양성과 각 소스간의 연계성 부족 문제를 해결하여야 하며, 이를 위한 방안으로 데이터 수집 메카니즘과 아울러 이들을 연결할 통합 환경이 매우 중요하다. 웹서비스기반 통합은 이러한 다양성에 대해 통합된 접근 방법을 제시할 수 있는 모범적인 대안이 될 수 있으리라 생각된다.

3. 결 론

이상으로 IT가 접목된 바이오인포매틱스 기반 기술을 소개하였다. **정보통신부에서의 바이오인포매틱스 기술 개발의 목표는 바이오산업을 첨단화하기 위하여 폭증하는 대용량 바이오 데이터를 분석하여 고부가 가치 정보로 가공하는 IT 접목 기술을 개발하는 것이다. 향후 바이오인포매틱스는 단순히 데이터의 처리, 분석, 저장, 관리 등의 실험실 보조 수단이 아니라 생물학 연구나 바이오산업에 필요한 제반 정보를 예측하는 고기능을 갖추는 방향으로 연구가 진행될 것이다.** 이를 위해서는 대용량의 데이터를 다각적으로 분석하여 나오는 지식을 종합하는 것이 매우 중요하다. 즉, 단편적인 지식이 아니라 입체적인 정보의 제공이 매우 필요하다고 하겠다. 또한바이오인포매틱스의 여러 요소기술들을 니드에 따라 워크플로우별로 통합하려는 이슈가 세계적으로도 제기되고 있다. 이는 사용자가 원하는 패키지를 맞춤형으로 제공하여 더 빨리, 더 간편하게, 더욱 정확한 정보를 제공하고자 하는 것이다. 사용자가 필요한 소프트웨어들을 워크플로우에 맞도록 조립하여 제공해 주는 통합 기술 또한 고려해야 하는 바이오인포매틱스의 중요한 분야가 된다. 결론적으로 **IT 시각에서 다루는 바이오인포매틱스의 주요 연구 주제는 고기능 정보처리 기술 개발과 개발된 요소 기술들의 통합이 될 것이다.**

참고문헌

[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 (PSB98).

[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.

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