대량 생물학 데이터의 모델링 및 마이닝을 통한 활용

이즈텍 이성근 · 김양석

1. 서 론

바야흐로 -omics 시대에 접어들면서 소규모 실험의 영역으로만 여겨지던 생물학에도 대량 분석의 길이 열리게 되었다. 인간 유전체 프로젝트(HGP: Human Genome Project)이후 유전체의 서열을 빠른 시간 내에 밝혀내는 방법들 덕분에, 생물학 문제는 근원적 성격을 띠는 1차원 DNA 서열 문제로의 전환이 대부분 가능하게 되었다. 이러한 서열 분석 위주의 유전체학(genomics)을 비롯하여 mRNA hybridization을 통해 유전자 발현 패턴을 봄으로써 많은 응용 가능성을 던져주고 있는 전사체(transcriptome) 분석, 단백질 전체의 상호 작용 네트워크 및 3차원 구조 분석을 통해 단백질의 기능과 기작을 연구하는 단백질체학(proteomics) 등 다양한 거시분석 분야가 새롭게 떠오르고 있다(그림 1 참조). 그렇다면, 소규모 실험실 단위의 기존 생물학 접근 방식이 지금까지 훌륭한 결과들을 내고 있음에도 불구하고 위의 방법들이 미래의 기술로 각광받는 이유는 무엇일까?

대체로 실험실 단위의 소규모 실험들은 in vivo 또는 in vitro 방법을 통해 생체과정의 한 부분을 국지적으로 집중하여 살펴봄으로써 거기에 대한 전문적인 지식을 얻게 되는 데는 더할 나위 없이 좋은 수단이다. 하지만, 하나의 생물체를 제대로 이해하기 위한 '진정한 해답'인 수많은 생물학적 개체(유전자, 단백질 등)간의 상호 작용 및 연관성 분석과 정량적 예측 모델링을 적용하기에는 한계가 있다. 이런 이유로 전통적인 생물학 실험과 분석 방법으로 얻어지는 결과들을 보완할 수 있고 이것들을 서로 연계시킬 수 있는 새로운 유형의 실험 시스템과 분석 기술이 필요한데, 계산생물학(Computational Biology) 혹은 생물정보학(bioinformatics)이 바로 이런 역할 수행의 중심에 서있는 것이다. 더구나 예전에는 개념적으로만 여겨지던 거시 분석 방법들이 'high-throughput'으로 명명 되어지는 새로운 기술들의 발전으로 인해 점차 현실화 되면서 생물정보학의 역할은 더욱 부각되고 있다. 이 글에서는 위의 거시 분석들 중에 필자들의 경험과 최근의 관심에 부합하는 DNA 칩과 전사체(transciptome) 분석, 단백질 상호작용 네트워크 분석에 연관된 방법들을 주로 논해 보고자 한다.

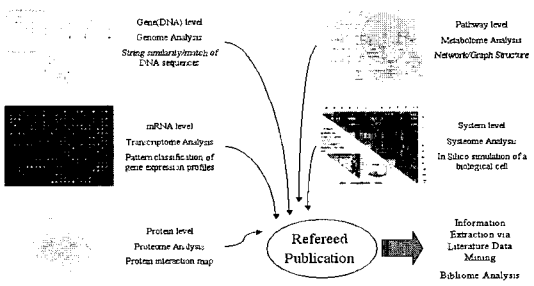


그림 1 다양한 형태의 -ome 분석 개요

2. DNA 칩 데이터 모델링 및 마이닝

DNA 칩은 조그마한 칩 위에 심겨진 수천 개 내지 수만 개 유전자들의 발현 양상을 어떤 실험조건들에 대해 동시에 관찰할 수 있다는 점에서 이른바 'snap-shot technology'라고도 부르고 있다. DNA 칩을 이용한 응용성은 매우 다양하여 유전자 기능 예측, 질병 진단, 독성 진단, 신약 개발에의 활용에 이르기까지 많은 잠재성을 지니고 있는 기술이다. 초창기에는 DNA 칩으로부터 산출되는 수치 데이터에 대한 데이터 정규화(normalization) 등의 전처리(preprocessing) 과정이나, 군집 분석(cluster analysis)등에 통계학자들의 많은 관심이 쏟아졌지만, 최근에는 판별분석(class prediction), 유의 유전자(differentially-

expressed genes) 선정 방법 등으로 연구 영역이 점차 확장되고 있으며, 이런 통계적 방법들의 신뢰도(significance) 문제도 꾸준히 연구되고 있다. 하지만, 이러한 문제들을 완전한 수학적 또는 전산학적 문제로만 바라본다면 큰 오산이다. 왜냐하면, 산출되는 결과를 생물학적으로 해석하는 작업이 반드시 수반되어야 하기 때문이다. 수학적으로 최적의 해가 반드시 생물학적으로 최적의 해가 된다고 말할 수는 없기 때문이다. 따라서 최근의 생물정보학 접근중의 하나는 이러한 수학적 전산학적 방법에 생물학적 지식을 결부시켜 정량적인 계산을 시도하는 것이다.

(주)이즈텍은 (주)SK와의 대형 GOM(Genomic Oriental Medicine) 프로젝트를 통해서 대량의 칩 분석을 수행하여 이미 다양한 분석 체계와 방법을 시행 경험한 바 있고, 현재에는 DNA 칩의 응용성에 중점을 둔 정부과제 3개를 수행하고 있으며, 대형병원, 대학 등의 외부 저명 연구기관과도 2-3 개의 프로젝트를 추진 중에 있다. 이외에도 집단유전학 관련 정부과제와 회사 내부 과제를 통해서, 칩 데이터에만 국한하지 않고 단일염기변이(SNP), 단백질 상호작용(Protein-Protein Interaction) 등 다른 데이터와의 상호 연관성에도 주목하고 있다. DNA 칩 데이터 분석은 언뜻 단순해 보이지만, 일반 통계 프로그램으로도 가능한 단순한 작업 이외에 칩 데이터 내에 감추어진 중요 정보를 제대로 마이닝하기 위해서는 정확하고 세밀한 유전자 주석 정보, 분석 방법의 통계적 검증, 타 생물학적 분석(ex: bio-pathway analysis)과의 연관성, 수치 분석 결과의 생물학적 해석 등 보다 전문화, 체계화된 절차와 방법이 요구된다고 하겠다. (주)이즈텍은 DNA 칩 데이터의 체계적인 통계 분석을 위해 S-plus 기반 소프트웨어인 CATMiD™(그림 2 참조)를 비롯하여 군집 분석의 통계적 수학적 해석 및 검증을 위한 ClusterVerifier™, 유전자, 단백질 등 생물학적 개체 그룹에 대한 최적 생물학적 의미 도출과 EST 등 대량 데이터의 생물학적 기능에 따른 분류가 가능한 GOODIES™ 등을 제작하여 DNA 칩 분석의 전문화 및 체계화를 갖추고 있다.

2.1 군집 분석(cluster analysis)

칩 분석 방법 중에서 가장 널리 알려진 군집 분석은 데이터 마이닝의 한 방법으로서 예전부터 많은 연구가 이루어지고 있는 분야이다. 결국은 수학적 고차원 공간상에서 주어진 데이터 개체를 그룹별로 묶는 것인데, 대부분의 경우 별다른 사전 지식(a priori knowledge)이 없는 채로 이루어지기 때문에 unsupervised classification 이라고도 한다. 최근에는 칩 데이터 군집 분석을 할 때, 유전자들 간의 이미 알려진 정보를 이용하여 supervised clustering을 하기도 한다[1].

각 개체를 그룹별로 묶을 때에는 어떤 기준이 있어야 하는데, 크게 두 가지로 개체 간 유사한 정도를 재는 척도와 사용자가 적용하고자 하는 모델에 따른 알고리즘이 있어야 한다. 유사성을 재는 척도로는 기하학적, 위상학적 고려를 감안하여 수학에서 정의된 다양한 거리(metric) 개념, 통계학에서 널리 쓰이는 여러가지 상관계수(correlation coefficient)와 정보학 분야에서 다루는 mutual information 등을 주로 쓰고 있다. 기본 모델 알고리즘은 [2]의 연구 이후에 많이 쓰이게 된 계층적 클러스터링(hierarchical clustering)을 비롯하여 K-means, PCA(Principal Component Analysis), SOM(Self-Organizing Maps), model-based clustering, graph clustering, biclustering에 이르기까지 실로 다양한 알고리즘들이 존재한다. 현재 추세를 보면, 저명 생물학 저널들에 발표되는 DNA 칩을 이용한 군집 분석에는 주로 계층적 클러스터링이 쓰이고 있는데, 이는 수형도(dendrogram)의 시각적 직관성이 주는 장점과 사용자가 원하는 군집을 패턴을 직

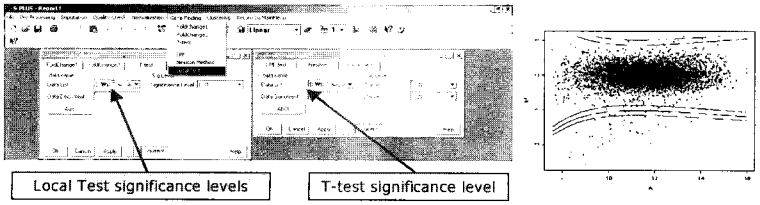


그림 2 CATMiD™: DNA 칩 데이터 통계 분석을 위한 통합 프로그램

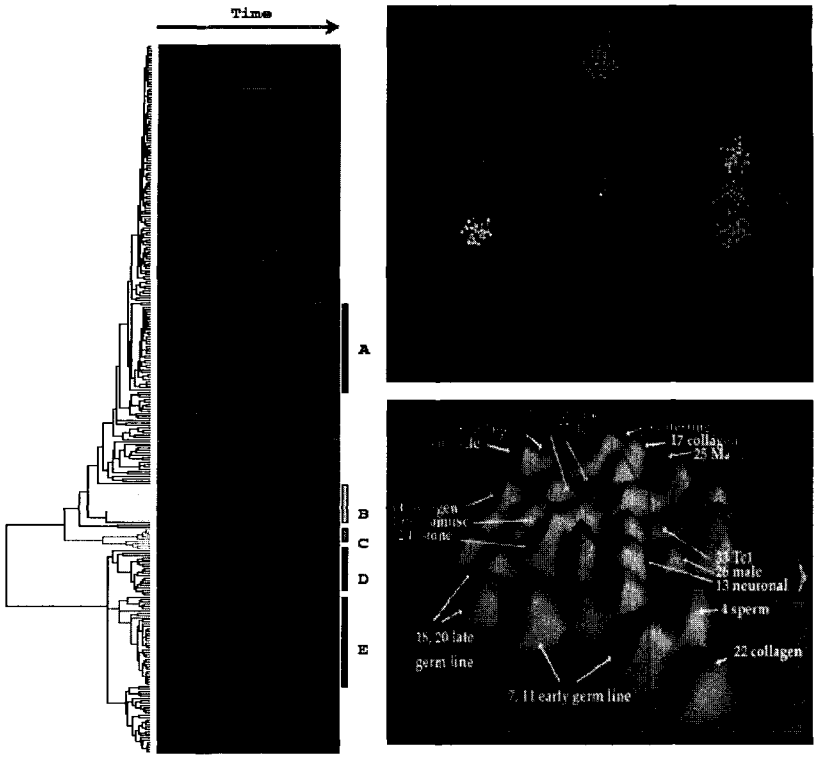


그림 3 군집 분석 결과의 시각화 : (왼쪽 그림) Eisen et al. (1998)의 수형도(dendrogram), (오른쪽 위 그림) PCA plotting from J. Quackenbush(2001), (오른쪽 아래 그림) topological 3D map by VxInsight

접 보면서 고를 수 있다는 측면에서 생물학자들에게 인기를 끄는 듯 하다. 이는 수치적으로 좋은 결과를 보이는 알고리즘이라 할지라도 실제 제품을 쓰는 사용자의 측면을 잘 고려하지 않으면, 단지 학문적인 관심으로만 남는다는 점을 일깨우는 것이다. 군집 분석 결과의 시각화는 이런 점에서 소프트웨어를 만드는 입장에서 주요 고려 대상이 된다(그림 3 참조).

칩 데이터를 분석함에 있어서 유전자에 대한 군집 분석의 기본 가정은, 생물학적 기능이 유사한 또는 동일한 조절 기작을 지니는 또는 동일한 생물학적 경로(bio-pathway) 상에 놓인 유전자들은 동일 시간상에서 유사한 발현 패턴을 보일 것이라는 가정이다. 물론 이런 가정은 time delay 등의 여러 복잡 다양한 생물학적 요인으로 항상 만족되지는 않지만, 지금까지도 여전히 유효하게 쓰이고 있다. 군집 분석의 응용 분야는 초기의 유전자 기능의 잠정적 예측으로부터 최근에는 의료적 응용성이 많이 부각되어 기존의 조직병리학으로 명확히 밝히기 힘든 암 종양들의 하위 그룹을 유전자 수준에서 구분할 수 있게 해주고 있다.

2.2 군집 분석의 검증(clustering validation)

여기서 검증이라는 의미는 군집 분석의 결과로 나온 군집들(clusters)이 과연 잘 묶였는지를 평가하고자 하는 것이다. 사실 다양한 알고리즘을 써서 군집 분석을 실행하였을 때 결과 해석을 어렵게 하는 점은 알고리즘마다 대체로 서로 다른 결과를 보인다는 것이다. 물론, 명확한 '경계'를 가지는 데이터의 경우에는 알고리즘간의 결과가 별다른 차이가 없는 경우도 있지만 이는 드문 경우이고, 일반적으로 알고리즘 간의 비교 뿐만 아니라 하나의 주어진 알고리즘에 대해서도 적절한 parameter 설정이 필요하다. 결국 최종적인 판단은 사용자의 몫이며, 이는 경험이 많고 숙달된 전문가가 아니면 쉽지 않은 작업이다.

이런 문제들을 해결하기 위해서 클러스터링 검증을 위한 여러 방법들이 제시되어 왔는데, 수치적인 접근과 생물학 데이터를 접목시킨 방법이 있다. 수치적 방법들은 아래의 질문들에 주목하고자 하는데, 즉, 주어진 데이터에 대해서 정확한 군집의 개수를 유추할 수 있는가? 군집의 개수를 구별하기 위한 명확한 군집의 경계는 어떻게 정의될 수 있을 것인가? 과연 그런 경계가 존재한다면, 수학적 고차원 공간상에서 어떤 계산을 통하여 찾아낼 것인가? 하는 궁극적인 질문들을 해결하고자 하는 것이다. 엄밀한 수학적 정의와 증명이 힘든 위와 같은 질문들을 정량적인 알고리즘을 통해 완전히 해결하기는 매우 어렵다. 수치적 검증의 하나의 해결책으로 누구나 납득할 수 있는 기본적인 아이디어는 다음과 같다. 군집 분석의 결과 중에서 각 클러스터 내부 거리가 최소화 되어 있고(compactness), 클러스터간 거리가 최대화 되어 있는(separation) 클러스터링이 있다면 그 군집 분할(partition)이 최적이 아닐까 하는 것이다. 이런 기초적인 아이디어를 발전시킨 여러 데이터 마이닝 index를 이용하여 군집 분석의 결과를 체계적으로 평가하는 방법들도 제시되었다[3, 4].

최근 군집 분석에서 주목할 점은 생물학 지식을 융합하는 분석 방법이다. 주된 두 가지 방법은 생물학적 온톨로지(ontology)와 주석 정보를 통한 접근과 문헌정보를 통한 접근 방식이 있다. 생물학적 온톨로지란 생물학적 어휘 분류 체계를 일컫는 것으로서 생물체에 연관된 각종 정보를 기술하는데 쓰이고 있으며, 특히 유전자나 단백질 등의 서열에 주석을 다는데 유용하게 쓰이고 있다. 대표적으로 Gene Ontology™, MIPS ontology, MeSH 등이 있으며, 특히 Gene Ontology™의 경우는 GO 국제 컨소시엄의 활발한 작업으로 꾸준하고 빠른 개선이 이루어지고 있다[5]. Gene Ontology™의 취지는 다양한 여러 생물체 종에서 제각각으로 쓰이고 있는 어휘 분류 체계들을 일관성을 갖춘 하나의 통합 어휘 체계로 제시하고자 하는 것이며 [6], 이런 시도는 비교 유전체학(comparative genomics)과 기능 유전체학(functional genomics)의 연구자들에게 도움을 주고 있다. (주)이즈텍의 GOODIES™는 유전자 주석 정보와 Gene Ontology™의 계층적 구조의 장점을 살린 그래프 모델링을 통해 군집 분석 결과의 생물학적 검증과 해석을 수행하고 있다(그림 4 참조). 생물학 문헌 정보를 통한 접근 방식은 [7] 등이 있는데, 각 유전자에 대한 정보를 MEDLINE 등의 텍스트 마이닝을 통해 유전자간 연관성을 정량적으로 계산하고 있다. 생물학 데이터베이스 내의 주석 정보는 업데이트 속도가 조금 느리지만 전문 생물학 주석자들의 손을 거친 정보이기 때문에 신뢰성이 높다는 장점이 있고, MEDLINE 등의 문헌정보를 통한 마이닝은 최신 결과가 잘 반영되지만 언어의 다양성과 불규칙성을 감안한 정확도 개선이 요구된다고 하겠다. 이런 여러 생물학 지식을 결부시킨 생물정보학적 접근은 앞으로 활발한 연구가 진행될 것이며, 기존의 수리적 분석방법들을 잘 보완해 줄 것이다.

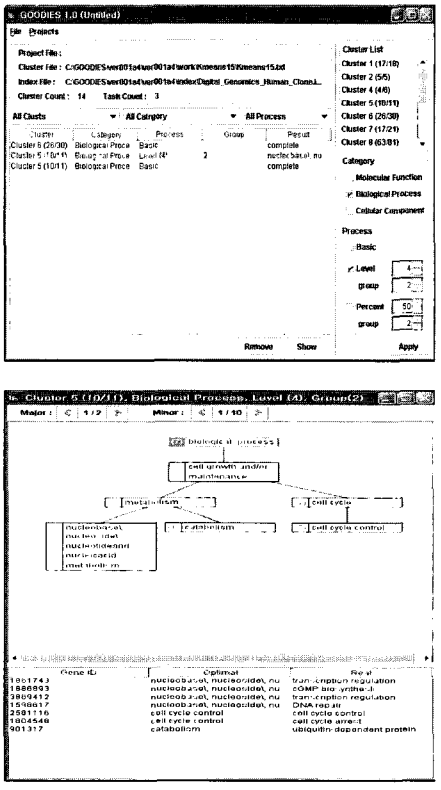


그림 4 GOODIES™: 대량 생물학 데이터 마이닝 프로그램(functional classification)

2.3 판별 분석(discrimination analysis)

판별 분석은 supervised classification이라고도 불리는 분야이다. 이미 알고 있는 사전 지식(training data)을 이용하여 입력 데이터가 소속될 그룹(class)

을 예측하는 것으로 최근에는 질병 진단, 암 종양 판별, 환자 생존기간 예측 등 임상적 측면들과 연계되어 많은 응용 잠재성을 지니고 있는 분야이다. 판별 분석은 세 가지의 관련 분야로 다시 나눌 수 있는데, 유의유전자 선정 또는 feature selection, 판별 알고리즘을 이용하여 판별기(classifier)를 만드는 부분, 판별기의 일반 오차율 또는 정확도를 예측하는 부분이다. (주)이즈텍은 2002년도에 식약청 과제 수행을 통해 만든 기본 판별 프로그램을 여러 공용데이터에 적용시켜 보았으며, 2003년도에는 보건복지부 과제로 질병 진단을 위한 좀더 발전된 전문 판별 소프트웨어를 제작 중에 있다. 현재 (주)이즈텍이 수행하고 있는 연구들 중에 이런 판별 분석과 연관된 것으로는 보건복지부 과제인 독성 진단 칩 개발 프로젝트와 백혈병 환자들의 특정 약물(글리벡)에 대한 민감도를 측정하기 위한 중소 기업청 프로젝트가 있다. 4년간의 장기 과제로서 추진 중인 독성 진단 칩 프로젝트는 여러 독성을 판별할 수 있는 독성 진단 칩을 개발하는 것으로서, 신약 개발의 전임상 단계에서 적용되는 동물실험의 대체 또는 감소 효과와 cDNA칩을 이용한 신약물질 검증의 효용성 증대에 그 목적이 있다. 데이터 분석을 맡고 있는 본사의 연구 방향은 칩에 심겨질 초기 유전자 선정을 위한 생물학적 데이터 마이닝, 각각의 독성에 연관된 유의 유전자 선정과 이를 통한 cDNA 진단 칩용 최종 유전자 선정, 기존 독성 관련 조직병리 데이터와 유전자 발현 칩 데이터와의 상호 연관성 및 이들 간의 정량적 관계, cDNA칩을 통한 독성 메커니즘의 규명 등인데, 성공적으로 프로젝트가 마무리되면 국내 신약개발에 큰 기여를 할 수 있을 것이라 예상된다.

2.4 칩 데이터 분석을 위한 기타 생물학 정보 마이닝

DNA 칩 실험의 결과로 나오는 것은 수치 데이터인지라 통계나 기타 수리적인 방법들이 칩 분석의 각 과정에 필수적으로 적용되지만, 최종적으로 생물학적 해석을 하기 위해서는 관련 생물학적 정보가 필요하다. 대체로 전문가들의 검토 과정(curation)을 거친 신뢰성 높은 생물학 데이터베이스들을 이용하게 되는데, 예를 들면, 다양한 통계적인 방법들로 유의 유전자들을 선정한 후에 이들 유전자들이 어느 생물학적 대사 경로에 관계되는지 알고 싶을 때는 KEGG, EcoCyc 등 생물학적 개체들의 알려진 기능 또는 관련 생물학적 과정의 주석 정보를 보기 위해서 MIPS, LocusLink, Swiss-Prot, TrEMBL, InterPro 등, 유전자들의 최종 생산물인 단백질들 간의 상호 작용을 알아보기 위해서 YPD, BIND, DIP 등 단백질 구조와 관련해서는 PDB 등을 검색하거나 필요 데이터를 추출하게 된다. (주)이즈텍은 각각의 특정 생물학적 정보에 대해 다양한 생물학 데이터베이스 간의 유기적 마이닝을 구축하고 있으며, 이를 통해 분석의 목적에 알맞게 재가공 하는 과정도 체계화, 자동화를 위해 꾸준한 개선 작업을 하고 있다. 이런 DB 마이닝 이외에도 기존에 나온 프리웨어들 중에 사용자의 요구에 부합되는 것이 있다면 유용하게 쓸 수 있을 텐데, 가령 [8] 등의 DNA 칩 발현 데이터-대사 경로 통합 시스템이나 기타 유전자 네트워크 모델링 등을 활용할 수 있을 것이다.

3. 단백질간의 상호작용체 분석 (interactome analysis)

단백질체(proteome) 분석 중에서 단백질 간의 상호작용(PPI: Protein-Protein Interaction) 연구는 최근에 그 중요성이 부쩍 높아졌다고 할 수 있다. PPI 분석은 단백질의 기능을 밝히는 데 도움을 줄 뿐만 아니라(functional proteomics) [9], 외부 환경 변화에 대한 반응 및 이에 대응하는 생체 시스템의 핵심인 신호전달체계(signal transduction pathway)의 여러 부분이 PPI로 이루어져 있어 이를 연구하는 것은 신약개발 사업에도 훌륭한 토대가 되는 것이다. (주)이즈텍은 PPI에 대해 다양한 생물정보학적 접근을 시도하고 있는데, 아래에 간략하게나마 언급하려 한다.

3.1 네트워크 측면에서의 검증

일반적으로 단백질체학(proteomics)에서 널리 쓰이고 있는 Y2H(Yeast Two Hybrid) 시스템이나 HMS(High-throughput Mass Spectrometry) 등에서 산출되는 대량 상호작용체(interactome) 데이터는 여러 장점에도 불구하고 비교적 높은 비율의 위양성(false positivity)을 지니고 있으며, 각 접근 방식에 따라 어느 정도의 데이터 편향성(bias)을 지닌다는 것이 알려져 있다[10]. 따라서 단백질 상호작용 데이터의 추출 못지않게 비교 및 검증 또한 중요한 문제

로 대두되고 있는데, (주)이즈텍의 MAPPI™는 여러 접근 방법들 중에서 그래프/네트워크 구조를 통한 접근 방식을 [11] 우선 시도하고 있다(그림 5 참조). 이런 다양한 방식으로 잡음 데이터를 제거하여 네트워크의 신뢰도를 개선함으로써, 양질의 단백질 상호 작용 데이터와 guilt-by-association의 적용을 통해 단백질 기능 유추 및 주석 작업에 큰 진전이 있으리라 예상된다.

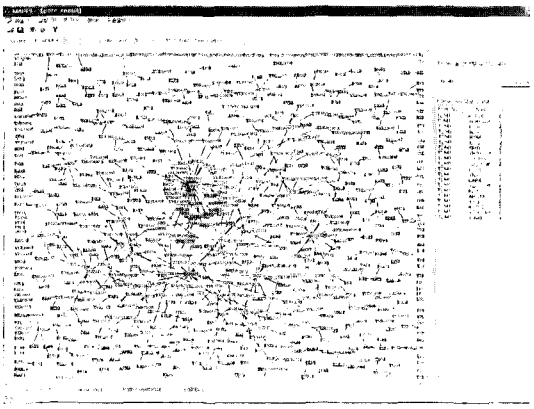


그림 5 MAPPI™: 단백질 상호작용 네트워크 분석 프로그램

3.2 전사체(transcriptome)와의 연관성

유전자들의 발현데이터 사이의 상관관계와 그 유전자들의 발현 결과로 생기는 단백질들 간의 상호작용은 과연 어떤 연관성이 있을까? 효모(S. cerevisiae) 등의 경우에 대해서 알려지고 있는 바는 위의 서로 다른 두 데이터 사이에는 통계적으로 - 또는 생물학적으로 - 어느 정도 의미 있는 관계를 보인다는 것이다 [12]. 물론 현재까지의 분석으로는 정량적으로 높은 상관관계를 보이는 것은 아니고, 대부분의 경우 임의의 유전자들 사이에서 벌어지는 상관관계보다 높다는 것이므로 이 부분에 대해서는 아직 많은 연구가 필요하다. [13]의 예처럼 유전자 발현 데이터의 클러스터링과 단백질체 상호작용을 서로 연관시킬 수도 있을 텐데, 이처럼 다른 실험 데이터들 간의 상호 연관성과 네트워크 관점에서의 이해는 앞으로 필수적일 것이다. (주)이즈텍의 차기 소프트웨어도 이런 점에 중점을 두고 개발될 예정이며, 이를 통해 기능 유전체학(functional genomics)이나 기능 단백질체학(functional proteomics)에 많은 도움이 될 것으로 기대한다.

4. 결 론

생물학이 정보 과학이나 수리 과학과 다른 점은, 지식을 기술하는 측면에서 정성적인 면(qualitative aspect)이 강하고, 생물체 내부 조절 기작과 신호 전달 경로가 상당히 복잡하고 동적이며, 생물 개체간 변이가 다양하고 그 변이 정도가 클 수 있다는 것이다. 계산생물학(computational biology) 또는 생물정보학은 정성적 관계 이면에 숨겨진 정량적 관계를 파악하려 하고 있으며, 이를 위해 선형 모델링으로 분해 단순화 시키거나 복잡계 시스템으로 해석하고 있다. 개체간의 변이나 대량 데이터에 내재된 불량 데이터를 효율적으로 처리하기 위해서는 필연적으로 통계적인 방법이 필요하고, 전산학 측면에서 기계 학습(machine learning) 기법들도 많이 적용되고 있는 추세이다. 하지만, 생물학적으로 복잡하고 어려운 문제는 수리적인 문제로 전환되어도 역시 어려운 경우가 흔히 있다. 가령, Motif 탐색이나 비교 유전체학(comparative genomics)에 요긴한 서열 분석 문제인 MSA(Multiple Sequence Alignment), 네트워크 비교 문제인 *subgraph isomorphism*, *protein folding* 모델링 관련 문제 등은 이론전산이나 수학분야의 최대 난제중 하나로 꼽히고 있는 NP-complete 문제들로 이미 알려져 있고, 군집 분석의 유용한 최적화 방법인 biclustering 관련 문제들은 NP-hard로 알려져 있는데, 앞으로의 생물학 데이터 마이닝이나 모델링 과정에서도 비슷한 상황들이 생길 수 있을 것이다. 이런 문제들은 '완전한' 해결책은 아직 찾지 못하고 있지만 비교적 좋은 결과를 주는 방법들이 속속 고안되고 있으므로 꾸준한 발전이 있을 것이라 믿는다.

생물정보학은 정보를 잘 저장하고 처리해서 생물학자들이 쓰기 쉽게 만들어주는 데이터베이스나 플랫폼의 성격으로만 인식될 수도 있다. 하지만, 이는 생물정보학 초창기의 협의적인 의미이며, 앞으로 다가올 생물정보학의 궁극적인 의미는 기존 생물학 연구의 커다란 패러다임 전환이 될 것이다. 인간 유전체 프로젝트 이후 장밋빛 미래에 부풀어 마치 우리가 당면한 모든 질병이 순식간에 해결될 것처럼 흥분한 때도 있었지만, 그런 흥분이 가라앉고 현재 우리 손에 쥐어진 것은 기존에 얻지 못한 엄청난 양의 데이

터이다. 이러한 대량 데이터 분석에 적절한 마이닝과 모델링을 동원하여 현재 당면한 문제들을 해결할 수 있다면, 맞춤 의학이나 복잡 다양한 질병들의 유전자 치료가 단지 꿈으로만 여겨지지는 않을 것이다. 쉬운 일은 아니겠지만, 지금껏 그래왔듯이 앞으로도 혁신적인 기술의 발전과 인간의 무한한 호기심은 우리를 희망에 찬 새로운 세계로 인도해 줄 거라고 믿는 바이다.

참고문헌

[1] M. Dettling & P. Buhlmann, "Supervised clustering of genes," Genome Biol. Vol.3, No.12, pp.research0069.1-0069.15, 2002.

[2] M.B. Eisen et al., "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci. USA, Vol.95, pp.14863-14868, 1998.

[3] F. Azuaje, "A cluster validity framework for genome expression data," Bioinformatics, Vol.18, pp.319-320, 2002.

[4] S.G. Lee et al., "ClusterVerifier: Statistical estimation tool for clustering validation of gene expression data of DNA microarrays.", Proceedings of the Annual Meeting of Korean Society for Bioinformatics, pp.277, 2002.

[5] Gene Ontology™ website: http://www.geneontology.org

[6] The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology," Nat. Genet., Vol.25, pp.25-29, 2000.

[7] Hovig et al., "A literature network of human genes for high-throughput analysis of gene expression," Nat. Genet., Vol.28, pp.21-28, 2001.

[8] B.R. Conklin et al. "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," Nature Genetics, Vol.31, pp.19-20, 2002.

[9] S. Oliver, "Guilt-by-association goes global," Nature, Vol.403, pp.601-603, 2000.

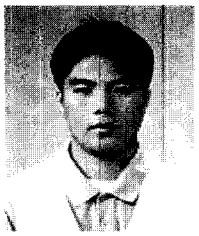
[10] von C. Mering et al., "Comparative assessment of large-scale data sets of protein-protein interactions," Nature, Vol.417, pp.399-403, 2002.

[11] R. Saito et al., "Interaction generality, a measurement to assess the reliability of a protein-protein interaction," Nucleic Acids Res., Vol.30, pp.1163-1168, 2002.

[12] A. Grigoriev, "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae," Nucleic Acids Res., Vol.29, pp.3513-3519, 2001.

[13] H. Ge et al., "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae," Nat. Genet., Vol.29, pp.482-486, 2001.

이 성 근



1996 포항공과대학교 수학과 석사
2002 포항공과대학교 수학과 박사
2001~2002 (주)이즈텍 객원 연구원
2002~현재 (주)이즈텍 선임 연구원(데이터마이닝 팀장)
E-mail : sglee@istech21.com

김 양 석

1998 포항공과대학교 생명과학과 박사
1999~2001 포항공과대학교 생물학 정보센터 부소장
2000~2002 National Cancer Institutue Biometirc and Bioinformatics Department visiting fellow
2002~현재 (주)이즈텍 대표이사
E-mail : yskim@istech21.com