단백질 상호작용 데이터베이스 현황 및 활용 방안

세종대학교 바이오인포매틱스 연구소

김 민 경·박 현 석

Protein Interaction Databases and Its Application

Min Kyung Kim and Hyun Seok Park

Institute of Bioinformatics, Sejong University, Seoul, Korea

ABSTRACT

In the past, bioinformatics was often regarded as a difficult and rather remote field, practiced only by computer scientists and not a practical tool available to biologists. However, the various on-going genome projects have had a serious impact on biological sciences in various ways and now there is little doubt that bioinformatics is an essential part of the research environment, with a wealth of biological information to analyze and predict. Fully sequenced genomes made us to have additional insights into the functional properties of the encoded proteins and made it possible to develop new tools and schemes for functional biology on a proteomic scale. Among those are the yeast two-hybrid system, mass spectrometry and microarray: the technology of choice to detect protein-protein interactions. These functional insights emerge as networks of interacting proteins, also known as "pathway informatics" or "interactomics". Without exception it is no longer possible to make advances in the signaling/regulatory pathway studies without integrating information technologies with experimental technologies. In this paper, we will introduce the databases of protein interaction worldwide and discuss several challenging issues regarding the actual implementation of databases. **(Immune Network 2002;2(3):125-132)**

Key Words: Protein interaction, bioinformatics, network

서 론

단백질의 기능을 규명함에 있어 그 단백질과 상호 작용을 하는 단백질을 찾아내거나 신호전달과정을 이해하는 것은 생물학자들의 주요 연구 방향 중 하나라 할 것이다. 이는 동일한 유전체 정보를 가지고 있는 세포들이 자신의 기능에 맞는 역할만을 수행하도록 특화되고, 외부 신호에 반응하며, 한 세포로부터 발생하여 개체를 이루는 등의 모든 생물학적 주제의 비밀을 푸는 핵심이었기 때문이다. 현재 단백질 상호 작용을 증명하기 위한 여러 실험 기법을 사용한 결과들이 보고되어 있다. 그렇다면 그러한 발견이나 실험을 하도록 제시해 주는 것은 무엇일까? 바로 그것은 이제까지의 결과, 즉 전산학적 용어로 말하자면 데이터베이스라 할 것이다. 새로운 염기서열이 등장하여 단백질로 추정되는 어떤 물질의 기능을 규명해야 한다면 가장 먼저 해보는 것이 무엇일까? 대부분의 연구자들이 주저 없이 염기서열분석을 시작할 것이다. 이제까지 알려진 사실들과의 비교를 통해서 어느 단백질과 유사성이 있는지, 혹은 어떤 기능을 암시하는 도메인, 모티프 등이 존재하는지 등의 사실을 바탕으로 하여 기능에 관한 가설을 세우고 그것을 검증하는 식의 접근은 궁극적으로 모든 생물학자가 진행하는 연구 방향이다. 하지만 실재 생물학자들이 서로 다른 정도와 수준에서 생물 데이터를 활용하고 있고, 그 중요성에 대해서도 서로 다른 견해를 가지고 있는 듯하다.

그럼에도 불구하고 앞으로 High Throughput Screening (HTS: 대량 분석 기술)으로 표현되는 실험들을 통하여

책임저자 : 김민경, 세종대학교 바이오인포매틱스 연구소
㉾ 143-747, 서울시 광진구 군자동 98번지
세종대학교 충무관 311A
Tel: 02-3408-3818, Fax: 02-3408-3881
E-mail: minkykim@sejong.mc.kr
www.sejong.ac.kr/~bioinfo, www.biopathway.or.kr

현재 세종대학교 바이오인포매틱스 연구소, 마크로젠, 스몰 소프트웨어, 서울대 그래픽스팀이 함께 농생물 생체 경로 지도 자동 구축용 툴 개발이라는 프로젝트(과제번호: AB-05)를 IMT-2000 출연금 기술 개발 지원 사업의 일환으로 농업과학 기술원의 연구 지원 아래 진행하고 있으며 이 기고는 연구 수행 과정 중 이루어진 기초 조사를 바탕으로 쓰여진 것임.

대량의 데이터가 점점 빠른 속도로 증가해 나갈 경향에 미루어 볼 때 *in silico*란 용어로 대변되는 dry lab에서의 분석 과정이 wet lab에서의 *in vivo, in vitro* 데이터를 도출해나가는 데 있어 중요한 비중을 차지할 것이다. 왜냐하면 양적으로 증가된 실험 결과를 처리함에 있어서도 생물정보학적 도구를 이용해야 할 뿐만 아니라, 데이터의 축적과 새로운 시스템의 개발 등에 힘입어 지금보다 더 정확하고, 수치적으로 표현된 예측이 가능해짐과 동시에 이러한 예측과 추론 과정을 통하여 생명 현상에 대한 좀더 포괄적이고 총체적인 이해를 할 수 있을 것으로 기대되기 때문이다.

현재 생물정보학은 단순한 서열 위주의 데이터베이스에서 상호작용을 주제로 하는 보다 더 고차원적인 데이터베이스화를 시도하고 있다. 이러한 데이터베이스들은 서열 및 구조 위주의 원시 데이터를 주내용으로 하는 데이터베이스들보다 특화된 주제를 다루고, 원시 데이터를 정밀하게 분석하여 새로운 패턴이나 조직화된 정보를 획득하는 것을 목적으로 한다. 따라서 연구자로서 실험을 계획하고 디자인함에 있어서 자신의 연구주제에 적합한 데이터베이스들을 이해하고 활용하는 지식을 습득해나가는 것이 필요할 것이다.

이 기고에서는 상호작용을 주제로 하는 데이터베이스와 그 과정에 등장하는 주요 이슈들에 대해 소개하고 연구자들의 관심과 활용 및 적절한 피드백을 통하여 생물학과 생물정보학의 상호 발전된 관계를 만들어 나가게 되기를 기대해본다.

본 문

생물정보학에서 상호 작용 데이터 베이스의 위치. 생물정보학은 인간 유전체 프로젝트와 같은 대단위 프로젝트의 등장과 DNA chip으로 대변되는 HTS의 등장으로 탄력을 받고 있는 것이 사실이다. 그러나 근본적으로는 생물정보학이 생명현상을 풀어나감에 있어 정보 처리의 관점에서 접근하고 있으며, 이것이 정보의 양이 기하급수적으로 증가하는 현 상황에 적합한 접근 방식이라는 것이다. 이전의 분자생물학적 접근 방식이 생명현상을 다룸에 있어 분자 수준에서 생명 현상을 설명하기 좋은 시스템이었다면 생물정보학은 유전체학, 단백질체학 등으로 불리는 총체적인 관점에서 생명현상을 이해하려는 움직임에 필연적으로 귀착된 방식임을 이해해야 한다.

실제로 컴퓨터와 인간의 생명현상이 서로 다른 정보처리 과정이 아니라는 사실이 여러 관점에서 입증되고 있다. 예를 들자면 컴퓨터가 0, 1을 정보의 단위로 사용하는 것과 마찬가지로 인간의 유전자는 A, T, G, C라는 4종류의 염기를 사용한다는 것이다. 실제로 전산학 알고리즘들 중에서 생명현상에 그 기본 아이디어를 가지고 있는 것이 많이 있는데, 유전자 알고리즘(genetic algorithm)이라든가 인공 뉴럴 넷(ANN: Artificial Neural Network), 인공 면역계(ARTIS: ARTificial Immune System) 등은 그 대표적인 것이라 할 수 있다.

생물정보학은 다루는 생물학적 정보의 내용에 따라서 (1) 서열을 다루는 유전체학-genomics, (2) 구조를 주관심사로 하는 구조 생물학-structural genomics, (3) 마이크로어레이나 DNA chip, 혹은 2D PAGE 등을 이용한 실험결과를 분석하여 발현을 연구하는 기능 유전체학, 혹은 단백질체학-functional genomics, proteomics, (4) 상호작용과 네트워크 등에 관심을 가지는 경로 생물정보학-pathway informatics, interactomics 등으로 구분될 수 있으며 단백질의 기능을 서열, 구조, 발현 및 상호작용의 관점에서 규명하여 총체적으로 이해하려는 공통된 목표를 가지고 있다.

서열이나 구조에 관한 생물정보학적 접근 방식의 역사에 비하여 단백질 상호작용에 대한 생물정보학적 연구는 90년대 후반에 들어서 본격화되었다고 할 수 있다. 서열이나 구조를 데이터로 표현하는 것이 보다 용이하고(서열은 한문자 알파벳으로 구조는 3차원 좌표값으로 표현된다), 인간 유전체 사업 등과 같은 대단위 유전체 사업들의 영향으로 그 데이터의 양이 급속히 증가하였기 때문이었다.

하지만 생물학자들에게 서열 그 자체보다는 상호작용이나 신호전달 과정에서의 그 단백질을 이해하는 것이 중요한 과제였음에 비추어볼 때 생물정보학에서도 상호작용을 주제로 하는 연구가 큰 비중을 차지하게 될 것임을 짐작할 수 있다.

단백질 상호작용 데이터베이스의 예들: 대사 경로의 경우도 단백질인 효소의 작용에 의해 주도되는 생체 내 반응이며 유전체 서열을 밝히는 작업 후에는 유전체 수준에서 대사경로를 구축하는 과제는 중요한 일임에 틀림없다(1-4). 실존하는 데이터베이스들 중 교토 대학의 KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.kegg.com)는 그러한 목적의 최초 시도로 인정받고 있으며, 현재 그 대상을 점점 넓혀 조절 반응 경로를 일부 포함하고 있다(5). 그러나, 이 논문에서는 주로 신호전달 경로를 이루는 단백질 대 단백질 상호작용에 초점을 두고 살펴볼 것이다.

BIND (Biomolecular INteraction Database, http://www.bind.ca); BIND는 캐나다의 토론토대학과 사무엘 루넨펠트 연구소의 공동 프로젝트로 진행되고 있는 사업의 일환으로 생체 내 분자들의 상호작용을 데이터화하는 작업을 진행하고 있다(6,7). BIND의 경우 현재의 주내용은 단백질 간의 상호작용이지만 단백질 이외의 DNA, RNA, ATP 등과 같은 small molecule이라는 범주로 포함시킨 물질들과의 상호작용까지 일반화시키고자 하는 의도를 가지고 있다. 상호작용 데이터의 표현 양식은 분자

A와 분자 B의 연결고리를 가지고 있는 interaction과 interaction이 서로 연결된 pathway, interaction의 다른 주체가 될 수 있는 complex 유형의 데이터를 지니고 있다. 따라서 분자 A 혹은 complex A와 상호작용을 하는 파트너의 리스트를 역동적으로 웹상에서 조사할 수 있고 그것이 지니고 있는 정보를 추출할 수 있다. BIND에서 특히 눈에 띄는 것 중의 하나는 preBIND라 명명된 기능으로 이는 문서로부터 경로 정보를 추출하며, 자연 언어처리(NLP: Natural Language Processing)라 불리는 전산학적인 기술을 사용한 것이다. Fig. 1에서 보는 것과 같이 상호작용을 포함하는 문서로부터 단백질 사이의 상호작용을 자동적으로 찾아주는 것으로 전문가의 리뷰를 통해 표시된 등급으로 데이터에 저장될 수도 있게 되어 있다. Fig. 1의 내용을 살펴보자면 ras1p를 검색어로 한 결과인데 Part 1에서는 ras1p 단백질의 기본 정보를 제공하며, Part 2에서는 ras1p와 상호작용에 관한 내용을 가지고 있을 가능성이 높은 논문의 리스트를, Part 3에서는 가능성이 적은 논문 리스트를 보여준다. 그중 하나의 논문을 선택하면 화면을 두 개로 분할하여 위에는 실제 초록 내용을 보여주고, 아래에는 그 논문에 등장하는 단백질 두 개를 테이블화하여 그 두 단백질이 상호작용이 실제로 그 논문에서 밝혀졌는가를 생물학적 지식이 있는 전문가의 검증을 받도록 되어있다. 따라서, BIND에서는 ras1p에 대한 검색 시 상호작용을 하는 파트너의 리스트와 함께 그것이 밝혀진 논문과 그 논문에서 사용된 실험기법 등의 정보까지 테이블로 보여준다. 그 결과들은 테이블 외에도 그래프로도 보여지는데, 자바 애플릿(JAVA applet)으로 구현된 프로그램을 통하여 하나의 단백질로부터 시작하여 그것과 상호작용을 하는 물질들을 차례로 열어갈 수 있다.

MINT (Molecular INTeraction database: http://cbm.bio.uniroma2.it/mint/; MINT는 로마대학(University of Rome Tor Vergata)에서 진행하고 있는 상호작용 데이터베이스이다(8). 이 사이트의 특징은 시각화와 데이터의 동기화(synchronization)이다. 따라서 그래프에서 해당 부위를 선택하는 것만으로도 연구자가 필요로 하는 정보로 이동하여 준다. 그 정보들은 전문가에 의해 정제된 것으로(생물학자들이 논문을 읽고, 필요한 자료들의 입력 과정을 거치는 되는데 이러한 역할을 수행하는 이를 curator라고 한다.), 상호작용이 일어나는 도메인 부위와 그를 증명한 실험, 각 단백질의 구조에 대한 데이터베이스 자료(PDB, Pfam, Prosite, SMART domain)를 가지고 있다. 사람에 의해 데이터를 축적해나가는 방식은 그 속도 면에서 한계가 있을 수밖에 없기 때문에 MINT의 경우에도 BIND와 마찬가지로 자연언어처리(MINT assistant라고 명명) 시스템을 염두에 두고 있다.

Fig. 2는 MINT 데이터베이스에서 p53과 상호작용을 하는 단백질의 네트워크를 살펴본 것이다. p53과 상호작용을 하는 것이 밝혀진 단백질로 MINT가 보유하고 있는 단백질의 개수는 16개이며, 그중 p73과의 연결선 위 물음표를 클릭하여 나타난 화면을 캡쳐한 것이다. Viewer에 의해 나타난 그래프 중 연결선 위의 물음표를 클릭함에 따라 테이블 내용도 두 단백질 사이의 상호작용에 대한 정보로 바뀌어 보여준다. p53과 p73은 그 결합이 co-immunoprecipitation의 방법으로 증명되었다는 것과 그 사실을 밝힌 논문 등의 정보를 얻을 수 있다. 그래프 표현에 대해 좀 더 자세히 설명하자면 각 단백질을 노드라 불리는 타원-분자량에 따라 크기를 정함-으로 표시하고, 원안의 작은 동그라미 혹은 선 위의 동그라미를 클릭하면 테이블의 내용은 그 단백질 혹은 상호작용에 관한 정보로 이동하며, 그래프는 그와 동시에 상호작용을 하는 모든 단백질을 펼쳐준다. 또한 각 단백질의 작은 동그라미 안에 +로 표시된 것은 그림에 나타나지 않은 상호작용이 존재한다는 것을 표시하는 전산학적 기호이다(하위 레벨에 존재하는 집합이 있다는 표시로 계층구조를 이루는 모든 개념들을 표시할 때 쓰인다). 검색어로는 단백질 이름뿐만 아니라 주제어(keyword)라든가, 서열 ID 등 다양한 방식으로 수행할 수 있도록 되어 있다.

DIP (Database of Interaction Proteins: http://dip.doe-mbi.ucla.edu/)와 proteinpathways (http://www.protein-pathways. com/); DIP은 UCLA의 아이젠버그 교수가 주도하고 있으며 몇 가지 점에서 주목할 필요가 있다(9). 노드(node)와 에지(edge)로 대변되는 데이터유형은 노드에 단백질을, 에지가 그 사이를 연결해주는 상호작용이라는 기본 개념을 가지고 있다. 노드에 해당하는 단백질에 대한 상세 정보에서는 그 단백질에 대한 cross reference 사이트를 링크로 걸어두고 있다. Visualization에 있어서 뷰는 그다지 좋지 않지만 그 의미는 상당히 중요한 의미를 지니고 있는데, 선의 굵기가 상호작용의 밝힌 실험적인 근거에 따라 신뢰도의 경중을 내포하고 있다. 따라서 생물학자들은 이 DB (databse)를 검색해보는 것만으로 어느 상호작용이 어떤 방법으로 증명되었는지 알 수 있으며, 이를 통해 더 높은 신뢰도의 실험을 진행하는 방식의 다음 실험 디자인에 대한 정보를 쉽게 얻을 수 있다(10). 이를 상업화한 시스템인 protein pathways에서는 논문으로 상호작용이 밝혀진 경우(Text Link: TL)뿐만 아니라 가능한 단백질 상호작용을 예측해 주는 시스템이 존재한다. Gene cluster (GC), phylogenetic profile (PP), gene neighbor (GN), Rosetta Stone (RS) 등의 정보를 활용하여 가능한 단백질 상호작용 파트너끼리 링크해주는 시스템은 아마도 인간의 일을 대신해주는 로보트에서 한 단계 더 나아가 지식적인 일을 해결해주는 know-bot의 원형이며 데이터를 push-down 해주는 시스

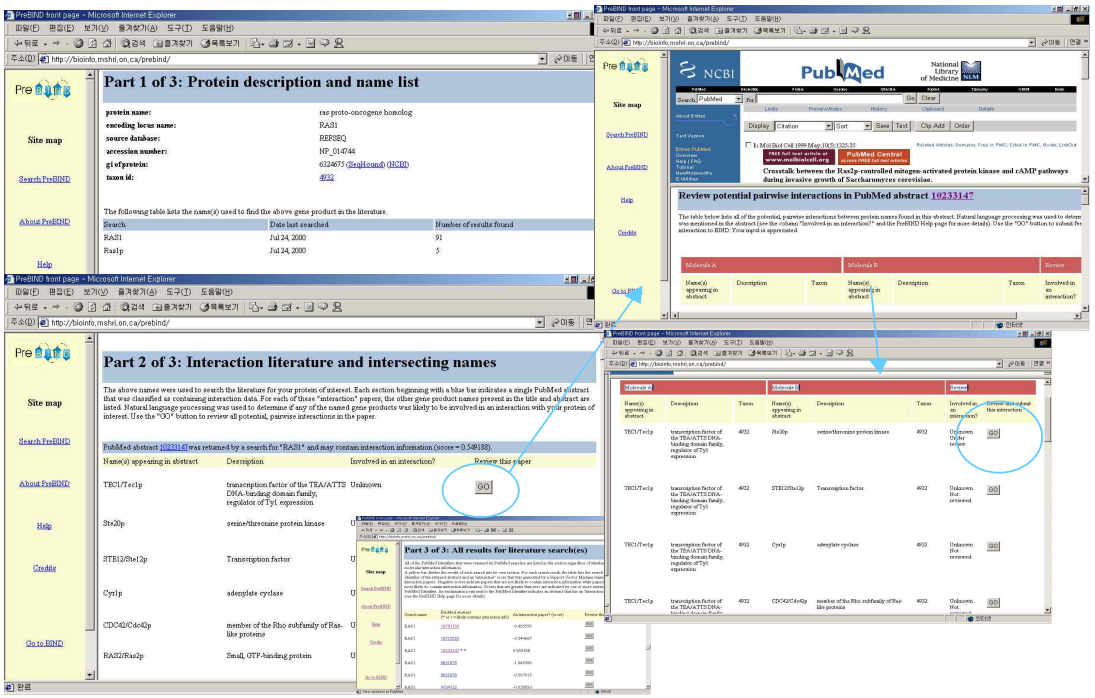

Figure 1. A search engine called preBIND, which is a support vector machine and NLP (natural language processing) based algorithm designed to identify abstracts that describe protein-protein interactions.

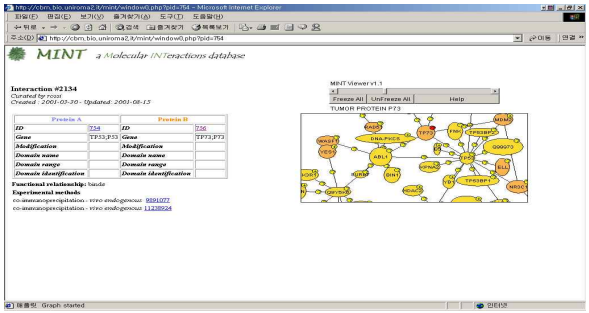

Figure 2. Database of MINT (Molecular INTeractions) which is devoted to the collection of protein-protein and protein-DNA interactions. New interaction data will be obtained from literature and will be provided through the collaboration of PhD students experimentally involved in interaction studies. Other data will be obtained from other interaction databases, when available. Tools for the graphic representation of interaction networks inside the cell will be developed. Each interaction will be recorded in association with the experimental techniques used to prove it.

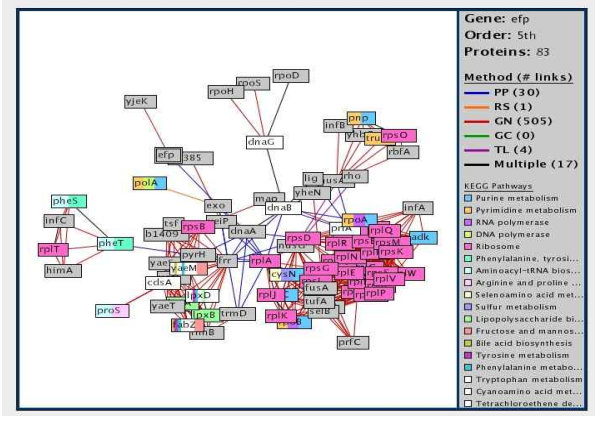

Figure 3. This is a demonstration which shows hands-on negotiation of Proteinpathway, Proteome Navigator platform. For this demonstration they have utilized E. coli version of ProLinks, specifically the links derived from the protein synthesis accessory proteins, EF-P. This input protein is linked to each other and to other proteins known to be involved in protein synthesis.

Table I. Protein interaction database collection

Service	URL	Characterist Ics
Kegg	http://www.kegg.com/	Support the curation of function assignments made to genes and the development of metabolic models
BIND	http://www.bind.ca/	Designed to store full descriptions of interactions, molecular complexes and pathways
DIP	http://dip.doe-mbi.ucla.edu/	Documents experimentally determined protein-protein interaction
MINT	http://cbm.bio.uniroma2.it/mint/	To store functional interactions between biological molecules (proteins, RNA, DNA).
MHCPEP	http://wehih.wehi.edu.au/mhcpep/	Database of MHC binding peptides
SYFPEITHI	http://syfpeithi.bmi-heidelberg.com/	Database of MHC ligands and peptide motifs
MHCBN	http://www.imtech.res.in/raghava/mhcbn/	Comprehensive database of MHC binding and non-binding peptides
FIMM	http://sdmc.krdl.org.sg:8080/fimm/	Integrated database of functional immunology, focusing on MHC, antigens, and diseases

템이 나올 것이라는 사실을 암시해주고 있다(11).

Fig. 3은 protein pathways에서 제공하는 가능한 단백질 상호작용을 그래프로 표시한 그림이다. 이 사이트는 공공 목적(public domain)이 아닌 상업화 시스템이기 때문에 몇 가지의 단백질에 대해서 데모만 돌아가며 Fig. 3은 EF-P (Elongation Factor-P)의 상호작용 파트너의 예측 결과이다. 따라서 가까운 장래에는 이러한 *in silico*상의 결과에 의해 수행할 실험의 범위를 제시받고, 그것을 *in vivo*, 혹은 *in vitro* 실험으로 확인해보는 경향이 더욱 증가될 것으로 생각된다.

MHCPEP (MHC-PEPtide binding database, http://wehih.wehi.edu.au/mhcpep/); MHC 분자와 상호작용을 하는 peptide에 대한 정보를 얻을 수 있는 곳으로 human, mouse, rat, chimpanzee, rhesus monkey, macaque, goat의 MHC class I, II에 결합하여 T 세포 활성화를 유발하는 13,000여 가지의 펩타이드 정보를 가지고 있다(12). 생물정보학자들은 이 DB를 이용하여 novel T cell epitope candidate peptides를 결정하는 예측 모델을 구축하는 것도 가능하고 이러한 예측모델을 사용하여 실제 생물학자들은 암 혹은 자가 면역 질환 연구 등에 활용이 가능할 것이다. SYFPEITHI (http://syfpeithi.bmi-heidelberg.com/), MHCBN (http://www.imtech.res.in/raghava/mhcbn/), FIMM (http://sdmc.krdl.org.sg:8080/fimm/) 등이 MHCPEP과 동일한 목적으로 만들어진 시스템들이며, 이외에도 여러 면역학 관련 연구자들에게 통합적인 정보를 제공하여주는 EBI의 IMGT (IMmunoGeneTics, http://www.ebi.ac.uk/imgt/) 등이 유용한 사이트가 될 것으로 보인다(13-15).

특히 면역학 관련 데이터베이스들은 면역에 관계된 단백질이 지닌 특유의 다형현상(polymorphism) 때문에 다른 주제들에 비하여 비교적 일찍 구축이 된 학문적 배경을 지니고 있다.

상호작용 데이터베이스의 주요 관심사.

문헌 정보 처리 기술: 단백질 상호작용에 관한 가장 많은 양의 정보를 가지고 있는 것 중의 하나는 아마도 문헌 데이터베이스일 것이다. 따라서 새로이 상호작용에 관한 데이터베이스를 구축함에 있어 전적으로 사람이 논문을 읽고 이해해서 데이터베이스화하는 방식에 의한 것이 아닌 컴퓨터의 자연언어처리(NLP: Natural Language Processing) 기술을 도입하고자 하는 시도가 있다. 이는 우선 상호작용에 관한 내용을 가지고 있는 문헌을 분류해서 문헌으로부터 상호작용을 나타내는 분자 A와 분자 B에 대한 내용을 끄집어내고 그것을 뒷받침하는 실험적 기술을 찾아주는 모양을 가지고 있다. 아직 자연언어처리 기술이 완벽하지 않기 때문에 사람의 검증 통한 확인 작업을 거치도록 하거나, 점수를 매기는 방식의 시스템이어야 한다. 상호작용이 있는 문헌을 분류하는 데에는 흔히 기계 학습 알고리즘(machine learning algorithm)이 적용된다. 이는 컴퓨터로 하여금 많은 양의 훈련 데이터로부터 상호작용을 나타내는 논문이 가지고 있는 규칙들을 발견하고 이를 실제 적용해서 시스템을 평가한다. 이는 실제로 사람이 논리적으로 찾을 수 없지만 훈련데이터 속에 내재해 있던 규칙들을 찾아줄 뿐만 아니라 대량의 데이터를 다룸에 있어 훨씬 유리한 측면이 있다. 이렇게 찾아진 문헌으로부터 상호작용을 의미하는 문장의 두 명사를-만약 대명사가 있다면 그를 처리할 수 있어야 하고, 부정어 등도 고려해야 한다-추출해서 관계를 지어주는 것이다(16-20).

Ontology 개념의 도입: 온톨로지의 개념은 데이터베이스 안에서 유전자의 기능 등을 서술함에 있어 보다 체계

적으로 정의된 어휘를 사용할 필요성에서부터 시작되었다. 생물학자들이 같은 의미를 다양한 방식으로 표현하고 이해하는 데 아무런 불편이 없지만 컴퓨터에서는 그러한 유형의 정의가 더 다루기 힘든 요소이다(실제 컴퓨터가 이해하는 프로그래밍 언어는 사람이 사용하는 언어보다 훨씬 더 정확한 규칙을 따르도록 정의된 문법의 일종이다). 따라서 일치된 어휘를 사용하고자 하는 이러한 노력은 정보를 공유하고, 재사용함에 있어 그 중요성을 가지게 된다고 할 수 있다. 특히 Gene Ontology 콘소시엄은 구조화되고, 잘 정의된 어휘로 유전자와 그 산물에 대한 개념화를 시도하는 것이다. 인간, 쥐, 초파리 등의 유전자에 대하여 최상위 cellular component, molecular function, biological process 3가지의 카테고리로부터 출발하여 그에 속하는 하위개념의 분류체계를 나무모양으로 가지치기를 하고 그 개념에 속하는 유전자를 위치시키는 방식이다. 어느 한 종류의 유전자 혹은 단백질은 그 기능이 여러 종류인 경우 혹은 보는 관점에 따라 여러 종류의 카테고리의 범주에 포함될 수 있다(21,22).

데이터 마이닝: 데이터마이닝이란 대규모의 데이터 내에 숨어 있는 고급 정보를 추출해서 예측, 예보 등에 응용하고자 하는 데이터베이스의 응용기술 분야이다. 생물정보학이 정보처리의 관점에서 데이터를 취급함에 있어 의미 있는 데이터의 패턴을 찾아내는 것은 생물정보학의 중요한 과제이다. 이는 인간이 더 의미 있는 것을 찾을 수 없다는 것이 아니라 시간이 너무 많이 들고, 번거로운 일을 단축시켜주기를 희망하는 것이다. 실제로 서열분석 등도 원하는 목표에 따라 시간을 빨리, 많은 양의 데이터와 비교해 보기 위하여 생물정보학적 접근을 하는 것이고, 그 정확성은 새로운 알고리즘의 도입, 데이터 양의 증가 등과 맞물려 눈에 띄게 개선되어 가고 있다. 따라서 무작위 대량의 데이터로부터 기계학습이라는 전산학적 기법을 사용하여 어떤 패턴이나 규칙, 새로운 정보를 알고자 하는 시도가 계속될 것으로 보인다.

상호작용 데이터베이스 등에서 기계학습을 통한 데이터 마이닝이 적용될 수 있는 분야는 상호작용 예측에 관한 것이다. 어느 정도 상호작용에 관한 데이터들이 축적이 되면 그 데이터들로부터 상호작용에 대한 일정한 규칙을 찾아내고, 알려지지 않은 상호작용을 예측할 수 있을 것이란 기대는 상호작용 데이터베이스를 구축하는 이들의 궁극적인 목표라고 할 수 있다.

이외에도 문헌정보에서 상호작용에 관한 논문을 찾음에 있어서도 똑같이 적용할 수 있다. 상호작용에 관한 논문이 가지고 있는 단어들의 유형을 발견하기를 희망하는데, BIND에서는 bind, interact, co-immunoprecipitation, yeast two hybrid 등의 용어를 포함하고 있는 경우 상호작용에 관한 내용을 가지고 있는 논문일 경우가 많다는 사실을 Support Vector Machine (SVM) 알고리즘으로 증명하여 사용하고 있다.

상호작용 데이터베이스와 HTS: 상호작용을 밝히는 주요 실험은 이제까지 yeast two-hybrid system으로 후보 물질을 찾은 다음 binding assay로 증명하는 방식으로 하나의 단백질에 대하여 그와 상호작용 단백질을 찾는 형태를 취해왔다. “omics”로 대변되는 synthetic approach의 경향성은 상호작용을 밝히는 것에까지 적용되어 유전체 단위에서 상호작용을 밝히고자 하는 시도가 있다. 이러한 시스템들은 bait로 하나의 유전자를 쓰는 것이 아니라 몇 백 개의 ORF를 사용하여 그 이전과 다른 스케일의 실험을 수행하는 것이다(23,24). 이러한 시스템들은 단순히 scale-up만을 목적으로 하는 것이 아니라 마이크로어레이 시스템 등과 같이 쓰여서 일정한 조건에서 특별히 상호작용을 하는 패스웨이를 찾거나 missing link를 찾는 목적에 유용하게 쓰일 수 있다(25,26). 마이크로어레이 데이터로부터 실재 내재한 패스웨이나 상호작용을 밝히는 것은 reverse engineering 작업이라고 할 수 있는데, 발현 데이터의 결과를 Boolean network, Petri Net 혹은 Graph theory 등을 라고 불리는 전산학 개념을 적용하여 상호작용을 밝히고자 하는 시도이다(27,28).

상호작용 데이터베이스의 활용 방안. 자신의 연구분야를 주제로 하고 있는 데이터베이스를 알고 활용하는 것은 실험 기술을 하나 알고 있는 것 못지 않게 여러 면에서 가치가 있다. 우선 비용 면에서 생각해보면 실제 실험을 하는 것보다 시간이나 노동력, 소모품비 등 여러 면에서 훨씬 절약할 수 있을 것이다. 또 다른 이점으로는 연구 범위를 좁혀 줄 수 있다. 이제까지의 밝혀진 사실을 토대로 어떤 사실을 새로이 밝혀야 하는지, 그 때 적용되어야 할 방법은 무엇인지, 내가 밝힌 사항이 이전의 다른 결과와 모순되지는 않는지 등등의 의문을 *in silico* 결과를 통해서 해결할 수 있다. 이러한 생물학자의 노력으로 밝혀진 결과들은 또다시 생물정보학자들의 데이터가 되고, 콘텐츠의 질은 전체 데이터베이스의 운명을 결정할 수도 있다. 그리고, 생물학자들은 자신들이 요구하는 기능이 무엇이며 무엇을 해결해주기를 기대해주는 지 등의 제안을 줄 수도 있다. 이러한 생물학자와 생물정보학자의 상호 교류가 서로에게 도움을 줄 수 있을 것으로 보인다.

고 찰

위에서 상호작용 데이터베이스들의 실제 예들과 그와 관련된 여러 이슈들에 대해 살펴보았다. 이러한 상호작용 데이터베이스들은 구축이 완성된 것이 아니라 현재 진행형인 사업들로써 밝혀진 모든 상호작용에 대한 내용을 담고 있지는 못하다. 실제로 이 사업은 단기간에 어느 독립적인 실험실이나 연구소의 노력에 의해 해결될 수 있는 주제가 아닌 것이 많이 존재하기 때문에 생

체경로 콘소시엄(BPC: The BioPathways Consortium)을 구성하여 표준화 등을 비롯한 기본 방향 등에 대하여 토의를 진행하고 있다(http://www.biopathways.org/). 최근 국내에서도 systems biology 관련 생물 정보 인프라 구축 작업이 학계와 정부 주도로 진행되고 있다.

지난 몇 십 년 동안 생물학자들의 연구 경향은 한 명의 학자가 하나의 유전자, 하나의 단백질 연구에 매달려 왔다면 앞으로의 연구 경향은 아마도 여러 유전자의 통합된 기능, 유전자 상호 작용 등에 관심을 가질 것이다. 지금 현재 NCBI에 등록된 유전체 서열이 800개 이상이라고 한다. 이는 서열의 유전체 사업이 얼마나 활발히 진행되고 있는지를 단적으로 보여주는 예이다. 또한 서열에서 더 나아가 유전체 단위에서 상호작용을 밝히고자 하는 시도가 이미 효모와 대장균 등의 경우에 논문으로 발표되었다. 다른 한편에서는 microarray, DNA chip 등의 기술을 이용하여 몇 만 개의 유전자 발현을 한꺼번에 보는 방법이 개발되어 있다. 따라서 앞으로는 유전자의 발현이나 상호작용을 통하여 단백질의 기능을 정확히 기술하고자 하는 시도가 더욱 중요한 일이 될 것임을 알 수 있다.

이러한 작업은 궁극적으로 어떤 자극의 결과를 예측하는 가상 시스템의 출현을 기대하고 있는데, 이러한 시뮬레이션의 가능성은 분화, 노화, 질병, 진화 등 모든 영역에 적용할 수 있다는 점에서 매력적이다. 실제로 이러한 움직임은 게이오 대학의 E-cell 프로젝트(http://www.e-cell .org/) 등을 통해 시도되고 있고 현재로서는 어느 정도 그 한계를 지니고 있지만, 지금의 발전 속도를 감안할 때 사람들의 생각보다 더 이른 시점에 이루어 질 수 있을 것으로 본다(29,30).

상호작용의 관점에서 생명현상을 이해하는 것의 장점은 서열과 구조와는 다른 시각을 제공하여 줄 것이라는 점이다. 서열 그 자체보다는 구조가 더 잘 보존되었다라는 사실에서 서열이라는 것이 구조를 만들기 위한 정보라는 측면을 이해할 수 있다면, 구조라는 것이 결국 상호작용을 위해 필요한 것이라는 관점은 어쩌면 구조보다 더 잘 보존된 상호작용의 어떤 새로운 패턴을 발견할 수 있다는 사실을 예견하고 있다.

감사의 글

세종대학교 바이오인포매틱스 연구소의 biopathway 프로젝트 자문 위원이신 캠브리지 대학교의 박종화 박사님과 OITEK의 오동훈 박사님, 연구에 대한 길을 열어주신 서울의대 박성회 교수님께 감사드립니다.

참 고 문 헌

1. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, Chemama Y, Labigne A, Legrain P: The protein-protein interaction map of *Helicobacter pylori*. Nature 409;211-215, 2001
2. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: Protein interaction maps for complete genomes based on gene fusion events. Nature 402;86-90, 1999
3. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: The large-scale organization of metabolic networks. Nature 407; 651-654, 2000
4. Paley SM, Karp PD: Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. Bioinformatics 18; 715-724, 2002
5. Kanehisa M, Goto S, Kawashima S, Nakaya A: The KEGG databases at GenomeNet. Nucleic Acids Res 30;42-46, 2002
6. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW: BIND-The Biomolecular Interaction Network Database. Nucleic Acids Res 29;242-245, 2001
7. Bader GD, Hogue CW: BIND-a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics 16;465-477, 2000
8. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: MINT: a Molecular INTeraction database. FEBS Lett 513;135-140, 2002
9. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30;303-305, 2002
10. Xenarios I, Eisenberg D: Protein interaction databases. Curr Opin Biotechnol 12;334-339, 2001
11. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: Detecting protein function and protein-protein interactions from genome sequences. Science 285;751-753, 1999
12. Brusic V, Rudy G, Harrison LC: MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res 26; 368-371, 1998
13. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50;213-219, 1999
14. Schonbach C, Koh JL, Flower DR, Wong L, Brusic V: FIMM, a database of functional molecular immunology: update 2002. Nucleic Acids Res 30;226-229, 2002
15. Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SG: IMGT/HLA Database-a sequence database for the human major histocompatibility complex. Nucleic Acids Res 29;210-213, 2001
16. Marcotte EM, Xenarios I, Eisenberg D: Mining literature for protein-protein interactions. Bioinformatics 17;359-63, 2001
17. Ono T, Hishigaki H, Tanigami A, Takagi T: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics 17;155-161, 2001
18. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M: Automatic extraction of protein interactions from scientific abstracts. Pac Symp Biocomput 5;538-549, 2000
19. Wong L: PIES, a protein interaction extraction system. Pac Symp Biocomput 6;520-531, 2001
20. Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J: Detecting gene relations from Medline abstracts. Pac Symp Biocomput 6;483-496, 2001
21. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G, Sethuraman A, Weng S, Botstein D, Cherry JM: Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids

Res 30;69-72, 2002
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. Nat Genet 25;25-29, 2000
23. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415;180-183, 2002
24. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403;623-627, 2000
25. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C: Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294; 2364-2368, 2001
26. Choi S, Hao W, Chen CK, Simon MI: Gene expression profiles of light-induced apoptosis in arrestin/rhodopsin kinase-deficient mouse retinas. Proc Natl Acad Sci 98;13096-13101, 2001
27. Liang S, Fuhrman S, Somogyi R: REVEAL, A general reverse engineering algorithm for inference of geneticnetwork architectures. Pac Symp Biocomput 3;18-29, 1998
28. Akutsu T, Miyano S, Kuhara S: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput 4;17-28, 1999
29. Tomita M: Whole cell simulation: A grand challenge of the 21st century. Trends Biotech 19;205-210, 2001
30. Tomita M, Hashimoto K, Takahashi K, Shimizu T, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, Hutchison C: E-CELL: Software environment for whole cell simulation. Bioinformatics 15;72-84, 1999