상호작용 및 도메인 정보를 이용한 단백질 기능 분석 시스템

김기봉
상명대학교 공과대학 생명정보공학과
kbkim@smu.ac.kr

Protein Function Analysis System Using Protein Interaction and Domain Information

Ki-Bong Kim
Dept. of Bioinformatics Engineering, Sangmyung University

요 약

기능 유전체학과 단백질체학에 있어서 개별 단백질의 기능 분석은 매우 중요한 핵심사안으로 대두되고 있다. 이러한 기능 분석에 있어서 과거와는 달리 현재는 전체 생명 시스템 상에서 개별 유전자 및 단백질의 기능 및 역할을 규명하는데 많은 초점을 맞추고 있다. 이러한 측면에서 단백질 상호작용 정보 및 도메인 정보를 기반으로 기능 분석을 행하는 것이 올바른 방법으로 인식되고 있으며, 본 논문에서는 그와 같은 분석 시스템을 소개하고 있다. 단백질 상호작용 정보는 모티프 및 도메인의 모듈 정보를 기반으로 하여 특이성과 민감도 측면에서 분석 정확성을 높일 수 있다.

1. 서 론

최근 엄청난 유전체 및 단백질체 정보들이 축적됨에 따라 독립 개체로서의 단백질 기능 보다는, 단백질 상호작용 네트워크의 구성요소로서 단백질 기능 규명에 연구의 초점이 맞추어지고 있다. 실용적인 측면에서는 특정 단백질간의 작용과 반작용은 신약 개발의 중요한 단서를 제공할 수 있다. 따라서 연구자들은 다양한 실험적인 방법과 전산학적 방법들을 사용하여 단백질간의 상호작용 관계를 밝혀내려 하고 있다[1]. 단백질의 상호작용 관계를 밝히기 위해 사용되는 실험적인 방법으로는 Yeast Two Hybrid 방법, Mass Spectrometry 방법, DNA Chip 방법 등이 있으며[1], 대표적인 계산적 방법으로는 계통발생 프로파일 방법, 유전자 인접 보존 방법, 유전자 융합 방법 등이 있다[1].

단백질 상호작용의 연구에서 중요하게 생각해야 할 부분은 단백질의 상호작용은 해당 단백질이 포함하고 있는 도메인 사이에 일어난다는 것이다. 하나의 단백질에 여러 개의 도메인이 포함된 경우, 실제로 상호작용이 일어나는 것은 하나 이상의 도메인들 간이다. 실제로 한 논문에서는 DIP (Database of Interaction Proteins)에 포함되어 있는 상호작용 단백질 쌍들에서 도메인 정보를 추출하고 상호작용할 가능성이 있는 PID (Potentially Interacting Domain)에서 TID (Truly Interacting Domain)를 가려내는 작업을 하였다[2]. 이런 점을 감안하여 본 논문에서도 단백질의 상호작용 관계를 도메인 수준에서 분석하고 그러한 상호작용 정보를 바탕으로 단백질의 기능을 추정할 수 있는 시스템을 소개하고자 한다.

현재 상호작용하는 단백질 쌍들에 대한 대표적인 데이터베이스들은 DIP (Database of Interacting Proteins), BIND (Biomolecular Interaction Network Database), GRID (General Repository for Interaction Datasets) 등이다. DIP은 단백질 상호작용 데이터베이스 중에서 가장 대표적인 것으로 현재 15,114개의 단백질 쌍들에 대한 데이터를 갖고 있다[3]. 그리고 BIND는 상호작용하는 단백질 쌍 이외에도, 분자 복합체 및 대사경로 등에 대한 정보들도 포함하고 있는데, 현재 11,237개의 단백질 쌍 엔트리를 포함하고 있다[4]. 마지막으로 GRID는 기존에 밝혀진 상호작용 단백질 쌍 데이터들을 통합하기 위해서 만든 데이터베이스로 20,984개의 상호작용 단백질 쌍에 대한 정보를 갖고 있으며, DIP과 BIND의 일부 데이터와 중복된다[5].

본 논문에서는 DIP, BIND, 그리고 GRID에 포함되어 있는 단백질 서열들을 도메인/모티프 통합 데이터베이스인 InterPro와 비교 분석하여 각 단백질의 도메인 정보를 추출하고, 그 정보를 바탕으로 도메인들 사이의 상호작용 관계를 밝혀내고자 하였다. 그리고 본 논문의 분석 시스템은 기존의 단백질 상호작용 데이터베이스들을 통합화하여 상호작용 단백질 쌍들에 대해 통합 검색 및 분석 기능을 제공하고 있으며, 서열 유사성과 도메인 수준에서 단백질의 상호작용 관계를 예측할 수 있도록 구성되어 있다.

2. 본 론

본 연구에서 개발한 기능 분석 시스템은 [그림 1]에서 볼 수 있듯이 기존의 상호작용 단백질 쌍들을 포함하고

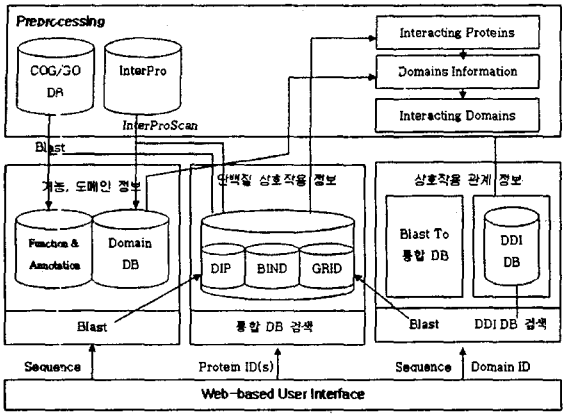

그림 1 전체 시스템의 구성도

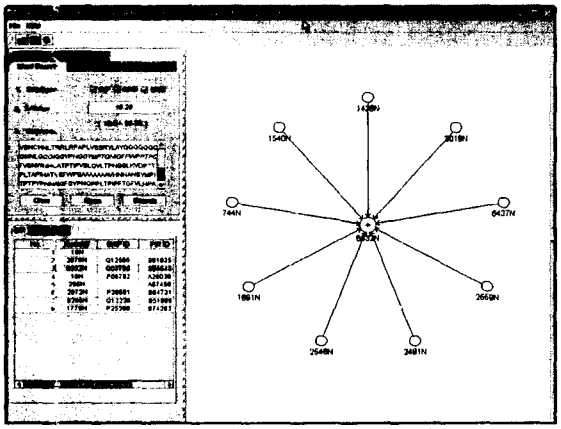

그림 2 분석 시스템의 메인화면

있는 데이터베이스들을 통합하였으며, 서열 유사성과 도메인 정보를 바탕으로 단백질의 상호작용 관계를 유추하고 가시화하여 사용자에게 보여주도록 구성되었다. 그리고 사전 작업으로 상호작용 단백질 쌍 데이터베이스에 포함된 서열들을 COG (Clusters of Orthologous Groups of proteins) 데이터베이스[6] 및 GO (Gene Ontology) 데이터베이스[7]에 대해 BLAST[8] 검색을 통해 단백질의 기능과 주석 정보들을 추출해 내고 그 결과를 분석 시스템 내에 데이터베이스화하였다. 따라서 사용자는 단백질의 상호작용 관계 유추뿐만 아니라 기능 및 주석 정보들을 한 번에 검색할 수 있게 시스템이 구성되어 있다. [그림 2]은 분석 시스템의 클라이언트 쪽의 메인 화면을 보여준다. 좌측의 상위 패널은 시스템의 주요메뉴 (DIP, BIND 및 GRID 데이터베이스 검색, BLAST 검색, Domain 검색)이며, 좌측 하위 패널은 좌측 상위 패널의 결과를 보여주는 것이다. 우측 패널은 단백질의 상호작용 관계를 가시화해서 보여주는 패널이다. 기능분석 시스템의 주요 기능을 살펴보면 다음과 같다.

2.1. DIP/BIND/GRID 의 통합 검색 기능

본 시스템은 기존의 데이터베이스들에 대해 통합 검색와 분석을 수행할 수 있도록 구성되어 있다. DIP 데이터베이스는 고유 번호를 가진 노드 (개개의 단백질)와 Edge (상호 작용하는 단백질 쌍) 데이터로 이루어져 있다. 각 노드는 SWP ID, GI ID, PIR ID 정보를 같이 저장하고 있으므로 사용자는 DIP 에 있는 단백질을 DIP 내의 노드 고유 번호 혹은 SWP ID, GI ID, PIR ID 로 검색할 수 있다. 검색한 결과에는 검색한 단백질의 서열, COG 로 나타나는 기능 정보, GO 로 표현되는 주석 정보, 그리고 그 단백질이 포함하고 있는 도메인 정보들이 포함된다. 그리고 우측 패널에서 검색 노드와 상호작용하는 단백질을 가시화해서 보여준다.

BIND 데이터베이스는 단백질의 고유 번호로 GI 를 사용하고 있으며 PubMed 의 Ref ID 정보와 명세를 함께 제공한다. 사용자는 GI ID, Ref ID 혹은 키워드로 BIND 에 포함되어 있는 단백질을 검색할 수 있다. 결과 패널에는 테이블이 나타나며 사용자가 키워드로 검색한 경우 여러 개의 단백질이 검색되며, 각 단백질을 클릭하면 그 단백질의 종, 서열, COG ID, GO ID 그리고 포함한 도메인 정보들을 확인할 수 있다. 그리고 DIP 과 마찬가지로 우측 패널에서는 상호작용하는 단백질들을 확인할 수 있다.

GRID 데이터베이스는 단백질의 고유 번호로 ORF ID 를 사용하고 있으며, Gene ID, SGD ID 를 함께 제공한다. 사용자는 ORF ID, Gene ID, SGD ID 로 GRID 데이터베이스를 검색할 수 있다. 결과 패널에는 DIP 검색에서와 마찬가지로 단백질의 서열, COG 로 나타나는 기능 정보, GO 로 표현되는 주석 정보, 그리고 그 단백질이 포함하고 있는 도메인 정보들이 포함되어 있다. 그리고 우측 패널에서 검색 노드와 상호작용하는 단백질들을 가시화하도록 구현되어 있다.

2.2. 단백질 상호작용 관계의 유추

단백질의 상호작용 관계를 유추하기 위해서 본 시스템에서는 유사성 기반과 도메인 기반의 2 가지 방법이 사용된다. 유사성 기반은 입력 단백질 서열을 DIP, BIND, 및 GRID 에 포함되어 있는 서열들과 BLAST 프로그램을 통한 상동성 검색을 통해 유사성이 발견되는 단백질 서열의 상호작용 관계를 바탕으로 입력 서열과 상호작용 할 것으로 추정되는 단백질들을 유추하는 방법이다. 이 방법은 단백질의 상호작용 관계를 단백질의 전체 서열에 의존하므로 실제로 상호작용이 일어나는 도메인 수준에서의 검토가 없다는 단점은 있지만, 일반적으로 단백질 상호작용의 유추를 위해 우선적으로 가장 많이 사용되는 방법이라 할 수 있다.

도메인 기반의 유추에 있어서는 두 가지 경우를 다룰 수 있게 하였다. 첫째는 사용자가 하나의 입력 도메인을 입력하는 경우이고, 둘째는 사용자가 두 개의 도메인을 입력하는 경우이다. 한 개의 도메인을 입력한 경우, 입력 도메인과 상호작용 하는 도메인의 정보를 보여주고, 두 개의 도메인을 입력한 경우에는 두 개의 도메인이 상호작용하는 빈도와 가능성 등을 보여주도록 구성하였다. 도메인 간의 상호작용에 대한 가능성은 아래 식의 계산값으로 수치화 하였다.

$$p(d_m, d_n) = \frac{1}{2}(1 + \frac{k_{mn}}{k_m k_n + \psi})$$

k_{mn} : number of edges in the training set
k_m : number of distinct vertices that contain at least one domain d_m
k_n : number of distinct vertices that contain at least one domain d_n
Ψ : pesudo-count (=1)

2.3. 상호작용 가능성이 가장 높은 단백질 쌍의 유추

생물학자가 여러 단백질 쌍들을 상호작용 할 후보로 예상하고 있을 때, 그 다중의 서열들을 본 연구의 분석 시스템에 입력하면 시스템 내부의 프로세싱을 통해 가장 상호작용할 가능성이 높은 하나의 단백질 쌍을 사용자에게 반환한다. 사용자가 입력한 각각의 단백질 서열은 InterProScan 을 통해 도메인 및 모티프 관련 데이터베이스들을 기반으로 도메인 영역의 검색이 이루지지고, 검색된 도메인 정보들을 바탕으로 도메인 기반의 단백질 상호작용 유추 방법을 이용하면 단백질의 상호작용 가능성을 수치화 할 수 있게 된다. 이것은 실험이나 분석 등을 통해 생물학자가 여러 개의 단백질 쌍들을 가지고 있는 경우, 효율적으로 이용될 수 있을 것이다.

2.4. 기타 기능들

앞에서 간략히 언급한 것처럼 분석 시스템 구축을 위해 전처리 작업으로 상호작용 단백질 쌍 데이터베이스에 포함되어 있는 단백질 서열을 COG 및 GO 데이터베이스와 BLAST 검색을 하여 단백질의 기능과 주석 정보들을 추출하고 분석 시스템 내에 데이터베이스화하였다. COG 와 GO 에 대해 BLAST 검색을 할 때, E-value 는 10^{-3} 으로 하였으며, 가장 유사성이 높은 것을 채택하였다. 따라서 사용자는 단백질 상호작용의 관계뿐만 아니라 입력 단백질의 기능 및 주석 정보들을 확인할 수 있다.

또 다른 사전 작업으로 상호작용 단백질 쌍 데이터베이스에 포함되어 있는 단백질 서열을 InterProScan 으로 InterPro 데이터베이스 (ProSite, ProDom, Prints, Pfam, Smart 등의 도메인/모티프 데이터베이스를 통합화 해 놓은 데이터베이스)와 비교하여 그 서열들의 도메인/모티프 정보를 추출하고, 그 결과를 데이터베이스화하였다. 따라서 사용자는 단백질 상호작용 관계뿐만 아니라 입력 단백질의 도메인 정보들을 확인할 수 있다.

3. 결 론

본 연구의 단백질 기능 분석 시스템은 대표적인 단백질 상호작용 데이터베이스들을 통합화하고 있다. 이러한 각 데이터베이스의 엔트리 서열에 대한 GO 및 COG 관련 유사성 정보와 InterPro 관련 모티프/도메인 정보 등을 전처리 과정을 통해 확보하여 시스템 내에 로컬 데이터베이스화하였다. 이렇게 시스템 내에 구축된 로컬 데이터베이스를 바탕으로 단백질 상호작용 관계를 추론함으로써 기능분석을 보다 체계적이고 효율적으로 행 할 수 있을 것이다. 즉, 사용자가 단백질의 상호작용 관계 유추뿐만 아니라 글로벌 수준의 상동성과 로컬 수준의 도메인/모티프 단계에서의 단백질 기능 및 주석 정보들을 총체적으로 획득할 있도록 시스템을 구현하였다. 특히 중요한 것은 서열기반의 단백질 기능 분석에 활용되는 GO 와 COG 데이터베이스들, 모티프/도메인 기반의 단백질 기능분석에 활용되는 InterPro 데이터베이스, 그리고 단백질 상호작용 기반의 단백질 기능 분석에 활용되는 대표적인 단백질 쌍 데이터베이스들인 DIP, BIND, 및 GRID 등을 총체적으로 활용하고 있어 분석의 효율성 뿐만 아니라 정확성까지 잘 반영하고 있다 할 수 있을 것이다.

4. 참고문헌

[1] Alfonso V., Florencio P., " Computational methods for the prediction of protein interactions" , *Current Opinion in Structural Biology*, Vol. 12, pp. 368-373, 2002.

[2] Wan K. K., Jong P., " Large Scale statistical prediction of protein-protein interaction by Potentially Interacting Domain(PID) pair" , *Genome Informatics*, Vol. 13, pp. 42-50, 2002.

[3] Loannis X. *et al.*, " DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions" , *Nucleic Acids Research*, Vol. 30, pp. 303-305, 2002.

[4] Gary B. *et al.*, " BIND : the Biomolecular Interaction Network Database" , *Nucleic Acids Research*, Vol. 31, pp. 248-250, 2003.

[5] Bobby-Joe B. *et al.*, " The GRID : The General Repository for Interaction Datasets" , *Genome Biology*, Vol. 4, pp 23-25, 2003.

[6] Tatusov *et al.*, " The COG database: new developments in phylogenetic classification of proteins from complete genomes" , *NAR*, Vol. 29, pp. 22-28, 2001.

[7] Steffen Hennig, Detlef Groth, and Hans Lehrach, " Automated 'GeneOntology annotation for anonymous sequence data" , NAR, Vol. 31, pp. 3712-3715, 2003.

[8] Altschul *et al.*, " Basic local alignment search tool" , *J. Mol. Biol.*, Vol. 215, pp. 403-410, 1990.