단백질 상호작용 추론 및 가시화 시스템

(A Visualization and Inference System for Protein-Protein Interaction)

이 미 경 † 김 기 봉 ††
(Mi-Kyung Lee) (Ki-Bong Kim)

요 약 다양한 유전체 프로젝트로 말미암아 엄청난 서열 데이타들이 쏟아지고, 이에 따라 핵산 및 단백질 수준의 서열 데이타 분석이 매우 중요하게 인식되고 있다. 특히 최근에는 단백질이 단순하게 개별적인 기능을 가진 독립적인 요소가 아닌 전체 단백질 상호작용 네트워크 상에서 다른 객체들과 유기적인 관계를 갖으며 나름대로의 중요한 역할을 수행하고 있다는 점에 초점을 맞추어 연구가 진행되고 있다. 특히 단백질 상호작용 관계 분석을 위해서는 실제로 상호작용이 일어나는 도메인 모듈 정보가 아주 중요하게 작용하는데, 본 논문에서는 이러한 점을 고려하여 우리가 개발한 단백질 기능 및 상호작용 분석을 위한 PIVS(Protein-protein interaction Inference and Visualization System)에 대해 소개하고 있다. PIVS는 기존의 단백질 상호작용 데이타베이스들을 합쳐서 통합 데이타베이스를 생성하고, 다양한 전처리 과정으로 통합 데이타베이스 서열의 기능 및 주석 정보를 추출하여 로컬 데이타베이스화 하였다. 그리고 특히 단백질 상호작용 관계 분석을 위해 중요하게 작용하는 도메인 모듈 정보들을 추출하여 로컬 데이타베이스를 구축하였고, 기존의 단백질 상호작용 관계 데이타를 이용하여 도메인 사이의 상호작용 관계 정보도 수집하여 분석하였다. PIVS는 단백질의 종합적인 분석 정보, 즉, 기능 및 주석, 도메인, 상호작용 관계 정보 등을 손쉽고 편리하게 얻고자 하는 사용자에게 매우 유용하게 사용될 수 있을 것이다.

키워드 : 유전체, 단백질 상호작용, 도메인, PIVS

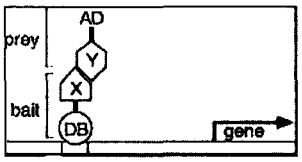
Abstract As various genome projects have produced enormous amount of biosequence data, functional sequence analysis in terms of the nucleic acid and protein becomes very significant. In functional genomics and proteomics, the functional analysis of each individual gene and protein remains a big challenge. Contrary to traditional studies, which regard proteins as not components of a whole protein interaction network but individual entities, recent studies have focused on examining functions and roles of each individual gene and protein in view of a whole life system. In this regard, it has been recognized as an appropriate method to analyze protein function on the basis of synthetic information of its interaction and domain modularity. In this context, this paper introduces the PIVS (Protein-protein interaction Inference & Visualization System), which predicts the interaction relationship of input proteins by taking advantage of information on homology degree, domain modules which input sequences contain, and protein interaction relationship. The information on domain modules can increase the accuracy of the function and interaction relationship analysis in terms of the specificity and sensitivity.

Key words : Genome, Protein interaction, Domain, PIVS

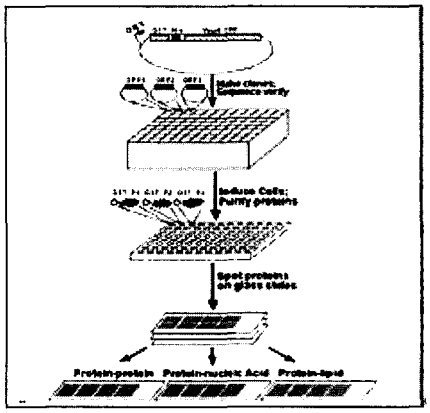
1. 서 론

최근 들어 다양한 유전체 서열이 밝혀짐에 따라 생명정보학적 측면에서 유전정보의 최종 산물인 해당 단백질을 밝혀 내고, 그 기능을 규명하는데 많은 연구가 진행되고 있다. 과거의 단백질 기능 분석 연구는 개별 단백질의 기능을 밝히는 데에 주안점을 두고 해당 연구가 진행되어 왔으나, 최근에는 대량의 유전체 정보가 밝혀짐에 따라 단백질은 하나의 독립적인 개별체로서 자신의 기능과 역할을 수행하는 것이 아니라, 생체내의 전체적인 단백질 상호작용 네트워크의 구성요소로서 기능과 역할을 수행한다는 점에 초점을 맞추고 있다. 그리고 실용적인 측면에서는 특정 단백질간의 작용과 반작용은 신약 개발의 중요한 단서를 제공할 수 있다. 따라서 단백질 기능 연구자들은 다양한 실험적인 방법과 전산학

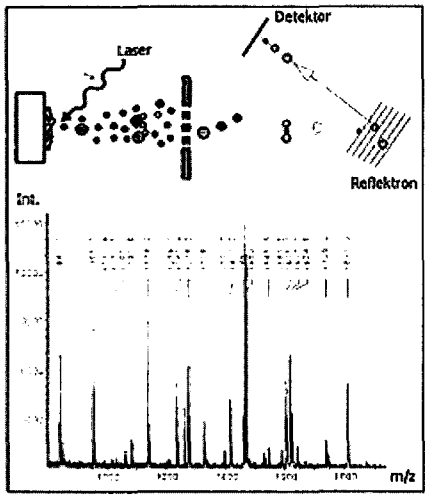
† 비 회 원 : 상명대학교 공학기술연구소 연구원
mklee@smu.ac.kr
†† 정 회 원 : 상명대학교 생명정보공학과 교수
kbkim@smu.ac.kr
논문접수 : 2004년 6월 1일
심사완료 : 2004년 9월 14일



(a) yeast two hybrid method



(b) Protein Chip



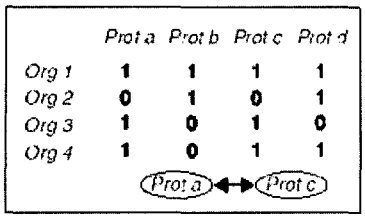
(c) Mass Spectrimetry

그림 1 단백질 상호작용 규명을 위한 실험적인 방법들

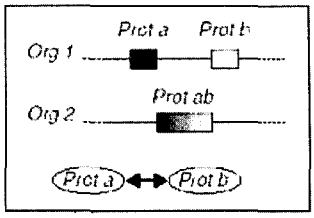
적 방법들을 사용하여 단백질간의 상호작용 관계를 밝혀내려고 하고 있다[1].

단백질의 상호작용 관계를 밝히기 위해 사용되는 실험적인 방법으로는 효모를 이용한 이종 하이브리드 방법(yeast two hybrid method), 질량 분광법(mass spectrometry), DNA 칩(DNA Chip) 및 단백질 칩(protein chip) 등이 있다(그림 1). 그러나 실험적인 방법은 결과의 정확성을 보장하기 어렵고, 한 번에 실험할 수 있는 양의 한계가 있으므로 전산학적인 방법을 병행하면 결과의 양적 및 질적인 측면을 향상시킬 수 있다.

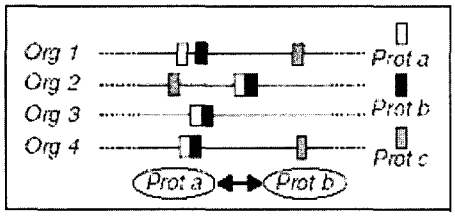
단백질의 상호작용관계를 밝히기 위해 사용되는 전산학적인 계산 방법으로는 계통발생프로파일 방법(phylogenetic profiles method), 인접유전자 보존성(conservation of gene neighborhood), 유전자 융합 방법(gene fusion method) 등이 있다[1]. 첫째, 계통발생프로파일 방법은 서로 상호작용하는 단백질은 진화론적으로 연관성을 띠는 여러 종들에서 같은 패턴으로 존재할 가능성이 높으므로, 각 단백질의 존재 유무를 프로파일링하고 그것의 유사성을 바탕으로 단백질의 상호작용 관계를 밝혀내는 것이다. 예를 들어 그림 2의 (a)에서 단백질(Prot로 표기된 것) a와 c는 유기체(Org로 표기된 것) 1, 3, 4에 똑같이 존재하는 것으로 보아서로 상호작용 하는 것으로 유추할 수 있다. 이 방법은 빠르고 계산양이 적다는 점에서 장점이 있지만 전체 유전체 정보가 밝혀져 있는 경우에만 적용될 수 있다는 점과 여러 생명체들의 생명 현상에 공통적인 즉, 기반적인 단백질들의 상호작용 관계를 밝히는 데에는 부적합하다. 둘째, 인접유전자 보존성은 서로 상호작용 하는 단백질들은 여러 종들에서 유전체 상의 물리적인 위치가 아주 가깝다는 것을 가정한 방법으로, 원핵생물 유전자들의 오페론 구조(operon structure)의 특성을 활용하고 있기 때문에 진핵 생물에는 적용될 수 없다는 단점이 있다. 예를 들면 그림 2의 (c)에서 단백질 a와 b는 유기체 1, 2, 3, 4 모두에서 물리적인 위치가 아주 가까이 존재하는 것으로 보아 동일한 기능이나 역할을 수행한다고 추정할 수 있고, 이에 따라 서로간에 상호작용 하는 것으로 볼 수 있다. 마지막으로 유전자융합 방법은 어떤 종에서는 개개의 독립적인 단백질로 존재하나 또 다른 종에서는 그 두 개의 단백질이 진화에 의해 합쳐져서 나타나는 경우이다. 이 방법은 대사체 단백질의 경우에만 국한해서 사용될 수 밖에 없고, 공유되는 도메인에 한정하여 유전자융합이 식별된다는 단점을 갖고 있다. 예를 들면 그림 2의 (b)에서 Org1에서 서로 독립적으로 존재하던 단백질 a와 b가 Org2에서는 합쳐져서 나타나므로 이들 단백질이 상호 작용할 것으로 예측할 수 있다. 그 밖에 단백질의 1차 구조와 관련된 정보 즉, 전하(charge), 소수성(hydrophobicity), 표면장력(surface tension) 정보들을 특징벡터(feature vector)로 표현하



(a) Phylogenetic Profiles



(b) Gene Fusion



(c) Conservation of gene neighborhood

그림 2 단백질 상호작용 규명을 위한 전산학적 방법들 (그림 출처 : 참고문헌 [1])

고, 그러한 특징벡터들을 SVM(Support Vector Machine)에 학습시켜서 단백질의 상호작용 관계를 유추하는 방법도 소개되었다[2]. 그리고 서열 유사성만을 바탕으로 단백질의 상호작용 관계를 유추할 때 생기는 false positive를 제거하기 위해서 그래프 군집화 방법을 도입하여 상호작용하는 단백질간의 군집을 만들고 군집 내의 도메인들 사이의 프로파일을 생성하는 시도도 이루어졌다[3]. 그 밖에, 데이타마이닝 기법의 하나인 연관 규칙 (association rule)을 이용해서 상호작용하는 단백질 쌍을 찾으려는 시도와 실제로 상호작용하는 데이타들을 효율적으로 보여주기 위한 다양한 가시화 방법들도 소개되었다[4,5].

단백질 상호작용의 연구에서 중요하게 생각해야 할 부분은 단백질은 하나 이상의 도메인들에 의해 모듈화된 구조를 띠고 있고, 이러한 도메인들이 다른 단백질의 도메인들과 상호작용함으로써 실제적인 단백질 상호작용이 일어난다는 점이다. 하나의 단백질에 여러 개의 도메인이 포함된 경우, 실제로 상호작용이 일어나는 것은 1개 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[6].

현재 상호작용하는 단백질 쌍들은 많은 사이트에서 데이타베이스 형태로 데이타를 제공하고 있는데, 대표적인 것으로 DIP(Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/), BIND(the Biomolecular Interaction Network Database, http://www.blueprint.org/bind/bind.php), GRID(the General Repository for Interaction Datasets, http://biodata.mshri.on.ca/grid/servlet/Index)가 있다. DIP은 단백질 상호작용 데이타 중에서 가장 많이 알려진 데이타베이스로써 15,114개의 단백질 쌍들이 포함되어 있다[7]. 그리고 BIND는 상호작용하는 단백질 쌍뿐만 아니라, 분자복합체(molecular complexes), 생체경로(pathways)에 관한 정보들도 포함하고 있는데, 11,237개의 단백질 쌍이 있다[8]. 마지막으로 GRID는 기존에 밝혀진 상호작용 단백질 쌍 데이타들을 통합하기 위해 만든 데이타베이스로 20,984개의 상호작용 단백질 쌍이 있으며, GRID의 데이타는 DIP과 BIND의 것과 일부 중복된다[9]. 그리고 2003년에는 InterDom이라는 데이타베이스가 소개 되었는데, 이것은 도메인 융합, 단백질 상호작용, 단백질 복합체 및 문헌정보를 이용해서 도메인간의 상호작용 정보를 수집하고 분석한 데이타베이스이다[10].

단백질 기능 분석을 위한 상호작용 관계 규명의 중요성을 감안하여, 본 논문에서는 다양한 수준, 즉, 서열들 사이의 전체적인 상동성, 서열들의 도메인 모듈 및 도메인-도메인 상호작용 수준 등에서 총체적으로 단백질 상호작용 관계를 분석할 수 있는 PIVS(Protein-protein interaction Inference and Visualization System)를 제안하고 있다. 특히 도메인-도메인 상호작용 수준에서 단백질의 상호작용 관계를 유추하기 위해서 우리는 DIP, BIND, 그리고 GRID에 포함되어 있는 엔트리 단백질 서열들을 도메인 및 모티프 관련 데이타베이스들을 통합 시켜놓은 InterPro 데이타베이스[11]와 비교하여 해

당 도메인 정보들을 추출하고, 이러한 정보를 바탕으로 도메인들 사이의 상호작용 관계를 밝혀내어 최종적으로 도메인 정보와 도메인-도메인 상호작용 정보를 로컬 데이타베이스화 하였다. 앞에서 소개한 InterDom의 경우는 단백질 상호작용 정보를 이용해서 도메인 간의 상호작용 정보를 밝히고자 할 때 PFam과 비교하여 도메인 정보를 추출하였지만, PIVS의 경우는 도메인 및 모티프 통합 데이타베이스인 InterPro(Prosite, ProDom, PRINTS, PFam, SMART, 및 TIGRFam 등을 통합시켜 놓은 것)를 이용하였으므로 분석 결과의 확장성과 정확성을 높일 수 있으리라 여겨진다. 그리고 본 시스템은 기존의 단백질 상호작용 데이타베이스들을 통합(X-Large DB라 명명함)하여, 사용자로 하여금 통합적으로 상호작용 단백질 쌍들을 검색 및 분석할 수 있도록 구성되었으며, 서열 유사성과 도메인 수준에서 단백질의 상호작용 관계를 예측할 수 있도록 구현되었다. 뿐만 아니라, PIVS 구축을 위한 전처리 과정의 일환으로 통합 데이타베이스인 X-Large DB에 포함된 각 서열들을 GO(Gene Ontology)[12]와 COG(Clusters of Orthologous Groups of proteins)[13] 데이타베이스와 유사성 검색을 수행하여 각 서열들의 주석 및 기능 정보 등을 뽑아내어 시스템내에 데이타베이스화 하였다. 따라서 사용자는 손쉽게 입력 서열의 종합적인 정보 즉, 도메인 정보, 주석 및 기능 정보, 그리고 상호작용 정보 등을 총체적으로 얻을 수 있다. 만약 상호작용 하는 두 개의 단백질 쌍이 같은 기능을 수행한다고 하면 그만큼 상호작용의 가능성은 커진다고 볼 수 있는데, 이와 같은 정보를 사용자는 아주 손쉽게 확인할 수 있게 되는 것이다. 다시 말해서, 기존의 단백질 상호작용 분석 시스템과 달리 PIVS는 단백질의 기능 및 주석 정보, 도메인 정보 및 상호작용 관계 정보 등을 총체적으로 제공하므로 일반 사용자가 본 시스템을 잘 활용하면 올바른 단백질 기능 및 상호작용 분석을 할 수 있을 것이다. 본 논문의 전체적인 구성은 다음과 같다. 2장에서는 PIVS의 전체적인 구성을 그림을 통해서 간략하게 설명하고 있으며, 3장에서는 PIVS가 일반 사용자에게 제공하는 기능들을 상세하게 기술하고 있다. 그리고 마지막 4장에서는 본 논문의 결론과 앞으로의 향후 연구과제 및 개선 방향 등을 제시하고자 한다.

2. PIVS의 전체 구성

앞에서 간략히 언급한 것처럼 PIVS는 전 세계적으로 가장 대표적이고, 많은 연구자들에 의해 널리 사용되는 상호작용 단백질 쌍 데이타베이스인 DIP, BIND 및 GRID 등을 통합화하여 자체적으로 하나의 상호작용 단백질 쌍 데이타베이스로 구축하였으며, 또한 단백질쌍 데이타베이스의 각 단백질 엔트리 서열들에 대해 서열 유사성 검색과 도메인 분석을 통해 얻어진 유사성 및 도메인 정보들을 시스템 내에 로컬 데이타베이스화하고, 이러한 정보를 바탕으로 단백질 상호작용 관계를 유추하

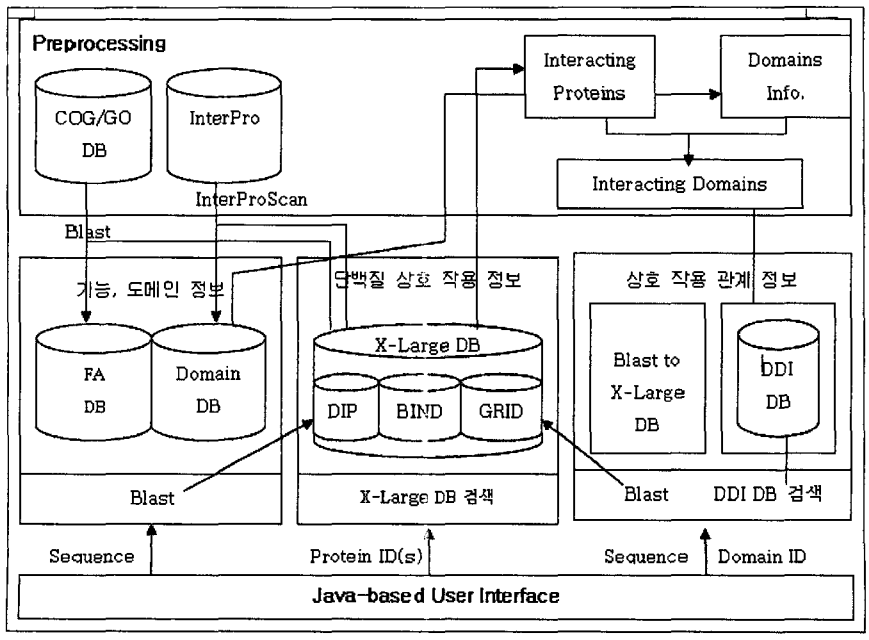


그림 3 PIVS의 전체 구성도

고 가시화하여 사용자에게 보여주도록 구성 및 구현하였다(그림 3). 보다 구체적으로 설명하자면 그림 3에서 볼 수 있듯이 기존의 데이타베이스들을 통합화하는 작업뿐만 아니라 여러 전처리 작업을 통해서 얻어진 데이타 및 정보들을 로컬 데이타베이스화 하여 사용자에게 총체적인 분석결과를 제공할 수 있도록 구성하였다. 전처리 작업은 크게 3개의 범주로 나누어 볼 수 있다. 첫째, 상호작용 단백질 쌍 데이타베이스에 포함되어 있는 단백질 엔트리 서열들을 COG(Clusters of Orthologous Groups of proteins)[12] 및 GO(Gene Ontology)[13] 데이타베이스를 대상으로 상동성 검색을 통해 단백질의 기능 및 주석(annotation) 정보들을 추출하여 내부적으로 FA (Function & Annotation) 데이타베이스를 구축하였다(그림 3 참조). 상동성 검색 시에는 대표적 상동성 검색 프로그램인 BLAST[14]를 사용하였고, 이때 사용한 E-value는 각각 10^{-3} 및 10^{-5}으로 하였으며, 가장 유사성이 높은 것을 채택하였다. 따라서 사용자는 단백질 상호작용 관계뿐만 아니라 입력 단백질의 기능 및 주석 정보들을 확인할 수 있다. 둘째, 상호작용 단백질 쌍 데이타베이스에 포함되어 있는 단백질 엔트리 서열들을 도메인 및 모티프 검색 프로그램인 InterProScan[15]으로 도메인 및 모티프 통합 데이타베이스인 InterPro[10]와 비교하여 각 단백질 엔트리 서열들의 도메인 및 모티프 정보를 추출하고, 그 결과를 이용하여 내부적으로 Domain 데이타베이스를 구축하였다(그림 3 참조). 따라서 사용자가 단백질 상호작용 관계뿐만 아니라 분석하고자 하는 입력 단백질의 도메인 정보들을 확인할 수 있게 시스템을 구성하였다. 셋째, 전처리 과정에서 얻어진 Domain 데이타베이스와 단백질 상호작용 쌍 통합 데이타베이스 내의 데이타를 이용해서 상호작용 단백질 쌍 정보와 해당 단백질들이 갖고 있는 도메인 정보를 이용하여 DDI (Domain-Domain Interaction) 데이타베이스를 구축하였다(그림 3 참조). 다시 말해서, 두 단백질이 상호작용하는 쌍으로 밝혀진 경우, 두 단백질이 포함하고 있는 도메인들은 실제로 상호작용할 가능성이 있는 도메인 쌍이라 할 수 있다. PIVS는 단백질의 상호작용 쌍과 그 단백질들이 포함하고 있는 도메인 정보를 이용하여 도메인 사이의 상호작용 쌍을 추출하고 이에 해당되는 정보를 데이타베이스화 하였다.

PIVS는 자바로 구현되어 플랫폼에 대해 독립적으로 사용될 수 있으며, 클라이언트, 서버 및 후미(back-end) 데이타베이스 등으로 구성되는 3-계층 구조를 이루고 있다. 서버는 BLAST 서버 모듈이 핵심을 이루고, 후미 데이타베이스 부분은 앞에서 언급한 단백질 상호작용관련 데이타베이스들과 자체적으로 전처리 작업을 통해 구축한 가공의 도메인 및 기능/주석정보 관련 데이타베이스들로 구성되며, 클라이언트는 GUI기반의 자바 응용 프로그램으로 이루어져 있다. 클라이언트 쪽에서 단백질 상호작용 관계를 가시화하기 위해 VCG (Visualization of Compiler Graphs) 패키지를 이용하였다[15].

3. PIVS의 주요 기능

3.1 기존 상호작용 단백질 쌍 데이타베이스들에 대한 통합 검색 기능

앞에서 언급한 것처럼 PIVS는 전세계적으로 일반 연구자들이 가장 널리 사용하는 상호작용 단백질 쌍 데이타베이스인 DIP, BIND 및 GRID 등을 통합화하여 자체적으로 하나의 상호작용 단백질 쌍 데이타베이스(즉, X-Large DB)로 구축하였다. 본 논문에서는 데이타 구조가 서로 다른 데이타들을 통합하기 위해서 X_LARGE_T 테이블을 생성하였다(그림 4 참조). 이 테이블은 inter_id, pro1_id, pro2_id, db_type 등으로 이루어진다. inter_id는 X_LARGE_T 테이블에서 사용하는 단백질 상호작용 쌍의 식별자이고, pro1_id와 pro2_id는 상호작용하는 두 단백질의 식별자를 나타낸다. 그 식별자는 DIP에서는 DIP 데이타베이스에서 사용하는 노드 아이디, BIND에서는 GI, GRID에서는 ORF 아이디이며 db_type은 해당 단백질 쌍의 데이타베이스 종류(즉, DIP, BIND, GRID 중 하나)를 표현한다. X_LARGE_T 테이블의 pro1_id와 pro2_id는 각 단백질들의 상세 정보를 저장하고 있는 DIP_NODE_T, BIND_NODE_T, 및 GRID_NODE_T 테이블 등의 해당 단백질 식별자를 참조한다(그림 4 참조). 사용자는 X-Large DB를 이용해서 손쉽게 여러 개의 단백질 상호작용 데이타베이스를 검색할 수 있다.

그 밖에 FA DB, Domain DB, 및 DDI DB 등을 위한 데이타베이스 스키마는 그림 5와 같은데, 기능 및 주석 정보는 COG_T 및 GO_T 테이블의 COG와 GO식별자를 통해 구별할 수 있으며 DOMAIN_T 테이블은 X-Large DB에 포함되어 있는 서열의 도메인 정보들을 모두 수집하여 저장한 테이블로서 도메인의 상세정보 및 위치정보 등을 가지고 있다. 그리고 DDI DB의 경우는 상호작용하는 도메인 쌍과 그 정보를 밝혀낸 상호작용하는 단백질 쌍 정보를 함께 저장하고 있다.

3.1.1 DIP 데이타베이스 검색 기능

DIP 데이타베이스는 고유 식별번호를 가진 노드(즉, 개별 단백질)와 에지(즉, 두 노드 단백질의 상호작용) 데이타로 이루어져 있다. 각 노드는 SwissProt ID (SWP), *Genbank* ID(GI), PIR(Protein Information Resource) ID 정보를 같이 저장하고 있으므로 사용자는 DIP에 있는 단백질을 DIP내의 노드 고유 식별번호 혹

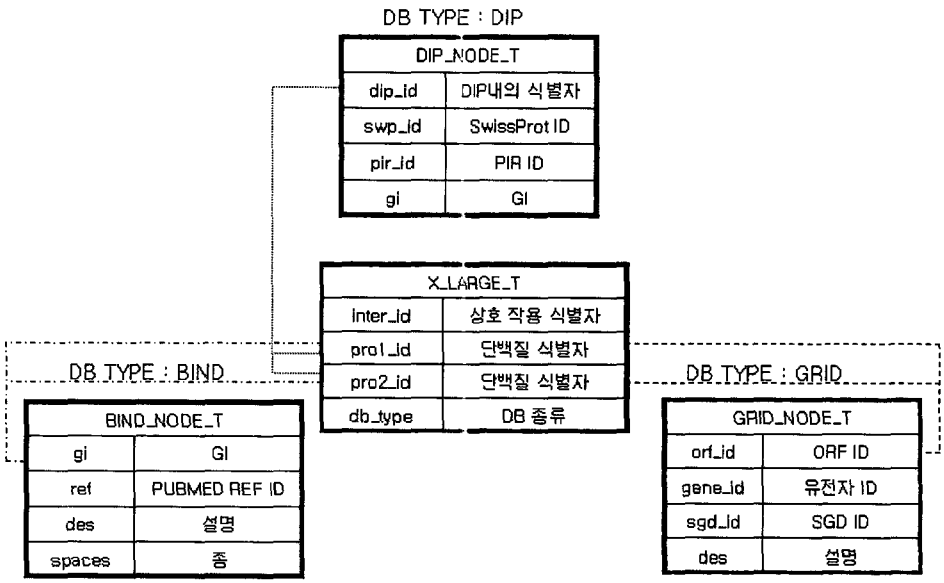


그림 4 X-Large DB의 데이타베이스 스키마

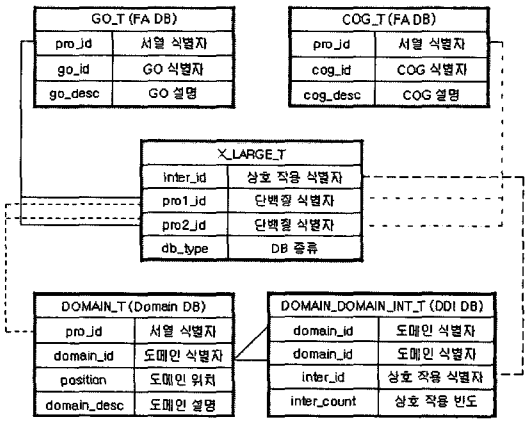


그림 5 단백질의 기능 및 주석, 도메인, 상호작용 정보를 위한 데이타베이스 스키마

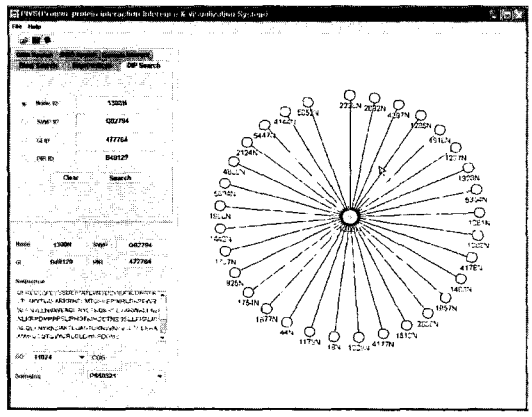


그림 6 DIP 데이타베이스 검색을 위한 옵션 설정 및 검색결과 화면

은 SWP ID, GI 및 PIR ID로 검색할 수 있다(그림 6 참조). 검색한 결과에는 검색한 단백질의 서열, COG로 나타나는 기능 정보, GO로 표현되는 주석 정보, 그리고 해당 단백질이 포함하고 있는 도메인 정보 등이 포함된다. 그리고 가시화 패널에서 검색 노드와 상호작용하는 단백질들을 그래픽하게 보여준다. 가시화 패널의 중심에 있는 노드가 검색한 노드이며, 검색 노드와 연결된 노드들은 모두 중심 노드와 상호작용하는 것들이다(그림 6 참조).

3.1.2 BIND 데이타베이스 검색 기능

BIND 데이타베이스는 단백질의 고유 식별번호로 GI를 사용하고 있으며 PubMed의 Ref ID 정보와 설명을 함께 제공한다. 사용자는 GI, Ref ID 혹은 키워드로 BIND에 포함되어 있는 단백질들을 검색할 수 있다. 검색결과 디스플레이 패널에는 테이블이 나타나며 사용자가 키워드로 검색한 경우 여러 개의 단백질이 찾아지는데, 각 단백질을 클릭하면 해당 단백질의 종, 서열, COG ID, GO ID 그리고 포함된 도메인 정보들을 확인할 수 있다. 그리고 DIP과 마찬가지로 상호작용 가시화 패널에서는 상호작용하는 단백질들을 확인할 수 있다.

3.1.3 GRID 데이타베이스 검색 기능

GRID 데이타베이스는 단백질의 고유 식별번호로 ORF(Open Reading Frame) ID를 사용하고 있으며,

Gene ID, SGD(Saccharomyces Genome Database) ID를 함께 제공한다. 사용자는 ORF ID, Gene ID, SGD ID로 GRID 데이타베이스를 검색할 수 있다. 검색 결과 디스플레이 패널에는 DIP 검색에서와 마찬가지로 단백질의 서열, COG로 나타나는 기능 정보, GO로 표현되는 주석정보, 그리고 해당 단백질이 포함하고 있는 도메인 정보들이 포함되어 있다. 그리고 상호작용 가시화 패널에는 검색 노드와 상호작용하는 단백질들을 보여준다.

3.2 단백질 상호작용 관계의 추론 기능

단백질의 상호작용 관계를 유추하기 위해서 PIVS에서는 유사성 기반과 도메인 기반의 2 가지 방법을 사용한다.

3.2.1 상동성 기반의 단백질 상호작용 관계 추론

상동성 기반은 입력 단백질 서열을 DIP, BIND 및 GRID 등에 포함되어 있는 서열들과 상동성 비교를 통하여 의미있는 유사성을 갖는 단백질 서열들의 상호작용 관계를 근거로 입력 서열과 상호작용할 것으로 유추되는 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. 그림 7의 좌측 상위 패널은 사용자가 단백질 서열을 입력하도록 구현하였으며 좌측 하위 패널은 사용자가 입력한 서열과 유사성을 보이는 서열들을 데이타베이스별로 보여주는 패널이다. 그리고 우측 패널은 좌측 하위 패널에서 선택한 단백질들의 상호작용 관계를 가시화해서 보여주도록 구현하였다.

3.2.2 도메인 기반의 단백질 상호작용 관계 추론

PIVS는 두 가지의 도메인 기반의 단백질 상호작용을 위한 인터페이스

추론방법을 제공한다. 첫째는 사용자가 하나의 입력 도메인을 입력하여 단백질 상호작용을 추론하는 방식이고, 둘째는 사용자가 두 개의 도메인을 입력하여 단백질 상호작용을 추론하는 방식이다. 첫 번째의 경우 사용자가 입력한 도메인과 상호작용하는 모든 도메인들의 정보를 보여주고, 두 번째의 경우는 사용자가 입력한 두 개의 도메인이 상호작용할 수 있는 빈도수와 가능성 등을 보여준다. 그 가능성은 식 (1)과 같이 계산된다. 식에서 k_{mn}은 d_m, d_n을 포함하는 두 단백질이 실제로 상호작용하는 횟수, k_m은 d_m을 포함하는 단백질의 수, k_n은 d_n을 포함하는 단백질의 수를 나타낸다. 그리고 사용자가 수치 확인을 쉽게 하기 위해 100을 곱하여 표현하였다.

$$p(d_m, d_n) = 100 \times \frac{k_{mn}}{k_m k_n} \quad (1)$$

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

그림 8은 도메인 기반의 단백질 상호작용 추론을 위한 인터페이스를 보여준다. 사용자가 두 개의 도메인을 입력하면 그 결과가 좌측 하위 패널에 나타난다. 입력한 두 도메인을 포함하고 있는 단백질의 수와 이들 도메인이 상호작용한 경우를 데이타베이스에서 추출하여 보여준다. 그리고 우측 윈도우는 입력한 두 도메인을 포함하는 단백질들의 목록을 보여주며, 두 도메인이 상호 용하는 경우를 데이타베이스 별로 보여준다. 그리고 사용자가 하나의 도메인을 입력하는 경우에는 입력 도메인과 상호작용하는 도메인과 그 빈도를 보여준다.

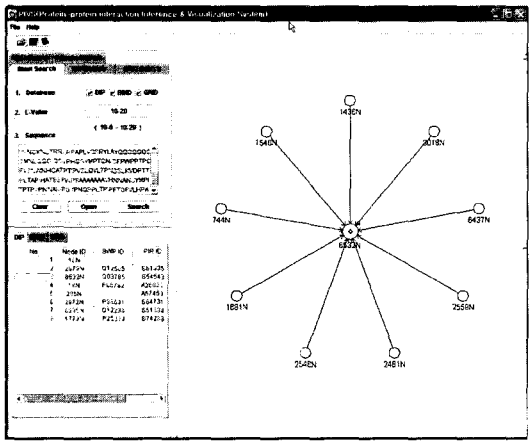


그림 7 상동성 검색을 통한 단백질 상호작용 관계 추론

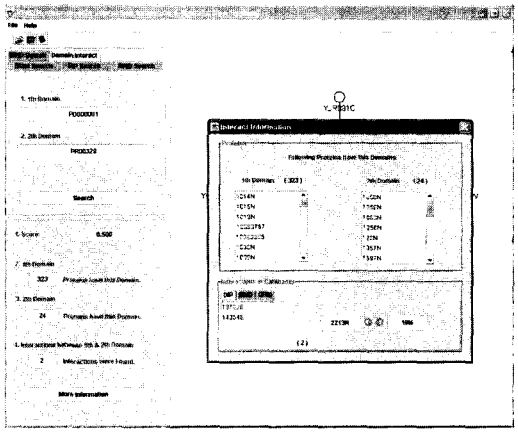


그림 8 도메인 기반의 단백질 상호작용 추론 인터페이스

표 1 상호작용 빈도가 50을, 점수가 3.0을 넘는 도메인 쌍들

No.	dm	dn	km	kn	kmn	p(dm, dn)
1	PF01423	PD020287	49	15	55	7.48299
2	SM00651	PD020287	49	15	55	7.48299
3	PF01423	PF01423	49	49	84	3.49854
4	PF01423	SM00651	49	49	84	3.49854
5	SM00651	SM00651	49	49	84	3.49854

4. 결론 및 향후 연구과제

4.1 결 론

본 논문은 단백질의 기능 분석을 위해 핵심적으로 요구되는 단백질의 상호작용 관계 및 기능 정보 등을 체계적으로 제공할 수 있는 PIVS에 대해서 언급하고 있다. PIVS는 기존의 단백질 기능 및 상호작용 관련 시스템과는 달리 분석하고자 하는 서열의 종합적인 정보 즉, 기능 및 주석 정보, 도메인 정보, 도메인 상호작용 정보, 단백질 상호작용 관계 정보 등을 일목요연하게 제공하므로, 효율적으로 단백질 기능 및 상호작용 분석을 하고자 하는 사용자들에게 매우 유익한 시스템이라 할 수 있다. 즉, 최종 사용자인 실험 연구자들이 정확한 기능 및 상호작용 분석을 위한 올바른 평가와 판단을 할 수 있는 체계적인 정보 및 데이타를 제공한다. PIVS의 기능과 특징은 다음과 같이 요약될 수 있다. 기존의 단백질 상호작용 데이타베이스들을 통합하여 사용자가 한번에 손쉽게 단백질 상호작용 데이타를 검색 및 분석 할 수 있다. 그리고 단백질 상호작용 데이타베이스의 ID나 서열의 상동성 검색을 통해 단백질의 상호작용 정보뿐만 아니라, 부가적으로 단백질의 기능 및 주석 정보, 그리고 단백질이 포함하고 있는 도메인 및 모티프 정보들을 확인할 수 있으므로 단백질 서열을 종합적으로 분석할 수 있다. 그리고 PIVS는 전처리 과정에서 단백질 수준에서의 상호작용 관계 정보를 이용해서 실제로 상호작용이 일어나는 도메인 수준에서의 상호작용 관계 정보를 데이타베이스화하여 사용자는 손쉽게 도메인 수준에서의 상호작용 정보를 추출할 수 있다.

본 논문에서는 도메인 수준에서 단백질의 상호작용 관계 및 기능을 규명하기 위해 다음과 같은 작업을 수행하였다. 우선 DIP, BIND 및 GRID 등에 있는 단백질들을 InterPro 데이타베이스에 대해 InterProScan으로 해당 단백질의 도메인 정보를 추출하였다. 그 결과 DIP 데이타베이스에 있는 단백질의 83% (5,534/6,652개)가 도메인을, BIND 데이타베이스는 76% (3,414/4,480개)가 도메인을, GRID 데이타베이스는 75% (4,417/5,888개)가 도메인을 포함하고 있었다. 위의 정보를 바탕으로 도메인 간의 상호작용 빈도 및 가능성을 추출하였는데, 384번으로 가장 높은 상호작용 빈도를 보인 도메인 쌍은 PS50322-PS00017인데, 해당 점수는 0.06591로 전체 상호작용 도메인 쌍 가운데에서 약 76%에 포함되는 수치이다. 그리고 상호작용 점수가 100.0으로 나온 PD000971-PS00540, PS00204-PS00540, PS00540-PS00540 세 쌍의 경우는 그 도메인을 포함하고 있는 단백질의 수가 모두 1이고 상호작용 횟수가 1로서 그 도메인이 나타나는 빈도 자체가 매우 낮으므로 신뢰할 만한 결과라고 보기는 어려울 것이다. 표 1은 상호작용 빈도가 50이 넘고 상호작용 점수가 3.0을 넘는 도메인 쌍을 보여준다. 본 논문에서 제시하는 도메인 상호작용 점수는 무작위 데이타의 분포를 고려하지 않고 단순히 상호작용이 보이는 도메인 쌍만을 가지고 계산하였으므로, 정확성이 떨어지는 단점이 있다. 그러나 실험 연구자들은 직관적으로 확인할 수 있는 도메인 상호작용의 빈도 및 그에 따른 점수를 확인할 수 있고, 이 정보뿐만 아니라 앞에서 언급했던 다른 여러 정보들을 취합해서 최종적인 판단을 내릴 수 있기 때문에 이러한 단점은 보완 및 극복될 수 있을 것이다.

그리고 본 논문에서는 서로 상호작용하는 단백질 쌍을 COG와 유사성 검색을 통해 두 단백질의 기능을 구명해 보았다. DIP은 2.1%(322/15,114)의, BIND는 1.6% (180/11230)의, GRID는 2.3%(493/20985)의 상호작용 단백질 쌍이 같은 기능을 수행하는 것으로 나타났다. 결론적으로, 기존의 단백질 상호작용 분석 시스템과 달리 PIVS는 단백질의 기능 및 주석, 도메인 정보, 상호작용 관계 정보 등을 총체적으로 제공하므로 사용자가 효율적이고 종합적으로 단백질 기능을 분석할 수 있다.

4.2 향후 연구과제

앞에서 언급했듯이 본 연구에서는 단백질 상호작용 정보, 기능 정보, 주석 정보 및 도메인 정보 등을 한 눈에 확인할 수 있는 PIVS를 구현하였다. PIVS는 단백질의 기능 분석을 위해서 단백질의 종합적인 정보를 얻고자 하는 사용자에게 큰 도움이 될 것으로 보이며, 앞으로는 도메인 수준에서의 상호작용 관계 정보를 바탕으로 보다 정확하고 신뢰할 만한 상호작용 도메인 쌍을 밝혀내도록 해야 할 것이다. 현재 PIVS는 자바 애플리케이션으로 개발되어 있는데 공공 서비스와 사용자 편의성의 양 측면을 고려한다면 웹 기반의 시스템으로 전환하는 것이 보다 나을 것으로 여겨진다. 그리고 상호작

용 관계를 추론함에 있어서 그 가능성을 좀 더 세밀하고 정확하게 측정할 수 있는 방법이 도입되어야 할 것이다. 뿐만 아니라, 갈수록 늘어나는 상호작용 데이타를 편리하게 시스템에 반영할 수 있는 자동 데이타 갱신 방법을 모색해야 할 것으로 여겨진다.

참 고 문 헌

[1] Valencia, A. and Pazos, F., "Computational methods for the prediction of protein interactions," *Current Opinion in Structural Biology*, Vol. 12, pp. 368-373, 2002.

[2] Bock, J. R. and Gough, A. D., "Predicting protein-protein interactions from primary structure," *Bioinformatics*, Vol. 17, No. 5, pp. 455-460, 2001.

[3] Wojcik, J. and Schachter, V., "Protein-protein interaction map inference using interacting domain profile pairs," *Bioinformatics*, Vol. 17, pp. S296-305, 2001.

[4] Oyama, T., Kitano, K., *et al.*, "Extraction of knowledge on protein-protein interaction by association rule discovery," *Bioinformatics*, Vol. 18, No. 5, pp. 705-714, 2002.

[5] Mrowka, R., "Java applet for visualizing protein-protein interaction," *Bioinformatics*, Vol. 17, No. 7, pp. 669-670, 2001.

[6] Kim, W. K., Park, J., and Suh, J. K., "Large Scale statistical prediction of protein-protein interaction by Potentially Interacting Domain(PID) pair," *Genome Informatics*, Vol. 13, pp. 42-50, 2002.

[7] Xenarios, L., Salwinsk, L., *et al.*, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, Vol. 30, No. 1, pp. 303-305, 2002.

[8] Bader, G.D., Donaldson, L., *et al.*, "BIND : the Biomolecular Interaction Network Database," *Nucleic Acids Research*, Vol. 31, No. 1, pp. 248-250, 2003.

[9] Breitkreutz, B.J., Stark, C., *et al.*, "The GRID : The General Repository for Interaction Datasets," *Genome Biology*, Vol. 4, R23, 2003.

[10] Ng, S.K., Zhang, Z., Tan, S.H., and Lin, K., "InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes," *Nucleic Acids Research*, Vol. 31, No. 1, pp. 251-254, 2003.

[11] Mulder, N.J., Apweiler, R., *et al.*, "The InterPro Database, 2003 brings increased coverage and new features," *Nucleic Acids Research*, Vol. 31, No. 1, pp. 315-318, 2003.

[12] Tatusov, R.L., Natale, D.A., *et al.*, "The COG database: new developments in phylogenetic classification of proteins from complete genomes," *Nucleic Acids Research*, Vol. 29, pp. 22-28, 2001.

[13] Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, Vol. 32, pp. D258-D261, 2004.

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

[15] Zdobnov, E. M. and Apweiler, R., "InterProScan - an integration platform for the signature- recognition methods in InterPro," *Bioinformatics*, Vol. 7, pp. 847-848, 2001.

[16] Visualization of Compiler Graphs - http://rw4.cs.uni-sb.de/users/sander/html/gsvcg1.html

이 미 경

2000년 부산대학교 독어독문학과 학사 2003년 부산대학교 대학원 전자 계산학과 석사. 2003년~현재 상명대학교 공학기술연구소. 관심분야는 바이오인포마틱스(단백질 구조, 단백질 상호 작용)

김 기 봉

2003년 9월~현재 상명대학교 생명정보공학과 전임강사. 1999년 5월~2003년 8월 ㈜스몰소프트 대표이사(실장/기술이사 역임). 1994년 3월~1999년 2월 한국과학기술연구원 생명공학연구소 연구원. 1998년 3월~2003년 8월 충남대학교 컴퓨터공학과 박사. 1995년 3월~1997년 2월 경북대학교 미생물학과 석사. 1985년 3월~1992년 2월 경북대학교 미생물학과 졸업. 관심분야는 생명정보학, 기계학습