Journal of Life Science **2004**, Vol. 14. No. 6. 997~1002 ©JLS

웹 기반의 단백질 상호작용 및 기능분석을 위한 보조 시스템 개발

정민철[1] 박 완[1] 김기봉*

상명대학교 공과대학 생명정보공학과, [1]경북대학교 자연과학대학 미생물학과,

Received September 21, 2004 / Accepted November 29, 2004

Development of Web-Based Assistant System for Protein-Protein Interaction and Function Analysis. Min-Chul Jung[1], Wan Park[1] and Ki-Bong Kim*. *Department of Bioinformatics Engineering, Sangmyung University, Chunan 330-720, Korea, [1]Department of Microbiology, Kyungpook National University, Daegu 702-701, Korea* – This paper deals with the WASPIFA (Web-based Assistant System for Protein-protein Interaction and Function Analysis) system that can provide the comprehensive information on protein-protein interaction and function concerned with function analysis. Different from existing systems for protein function and protein-protein interaction analysis, which provide fragmentary information restricted to specific field, our system furnishes end-user with comprehensive and synthetic information on the input sequence to be analyzed, including function and annotation information, domain information, and interaction relationship information. The synthetic information that our system contains as local databases has been extracted from many resources related to function, annotation, motif and domain by various pre-processing. Employing our system, end-users can evaluate and judge the synthetic results to do protein interaction and function analysis effectively. In addition, the WASPIFA system is equipped with automatic system management and data update function that facilitates system manager to maintain and manage it efficiently.

Key words – Protein-protein interaction, WASPIFA system, domain, pre-processing, annotation information, domain, local database, automatic data update.

다양한 유전체 염기 서열들이 밝혀짐에 따라 이들이 암호화하고 있는 단백질의 기능을 전산학적 기법으로 규명하고자 하는 연구가 시스템 생물학적 측면에서 많이 진행되고 있다. 과거의 단백질 기능분석은 일반적으로 각 개별 단백질의 기능을 밝히는 데에 초점을 두고 연구가 진행되었다. 그러나 최근에는 개별 유기체의 유전체 전체의 서열정보가 밝혀짐에 따라서 단백질이 하나의 독립체로서 해당 기능을 수행하는 것이 아니라, 전체적인 단백질 상호작용 네트워크 및 조절 네트워크의 구성인자로서 유기적인 네트워킹을 통해 해당 기능을 수행한다는 점에 주안점을 두고 있다. 다른 한편으로, 실용적인 측면에서는 특정 단백질간의 작용과 반작용은 신약개발의 중요한 단서를 제공한다. 따라서 연구자들은 다양한 실험적 방법과 계산학적 방법 등을 이용하여 단백질 상호작용 관계를 규명하고자 한다[3,8]. 질량 분광법(Mass spectrometry), DNA 및 단백질 칩, Yeast Two Hybrid 방법 등이 대표적인 실험적 방법으로 대량의 실험 데이터를 양산하고 있고, 전산학적 방법들로는 계통발생 프로파일(Phylogenetic profiles) 방법, 유전자 인접보존(Conservation of gene neighborhood) 방법, 유전자 융합(Gene fusion) 방법 등이 널리 사용되고 있다[8]. 이외에도 단백질의 1차 구조와 관련된 정보 즉, 전하, 소수성, 표면장력 정보 등을 특징 벡터(feature vector) 값으로 하여 SVM (Support Vector Machine)을 학습하고, 그러한 학습을 바탕으로 단백질의 상호작용 관계를 유추하는 방법도 소개되었다[3].

***Corresponding author**
Tel : +82-41-550-5377, Fax : +82-41-550-5184
E-mail : kbkim@smu.ac.kr

단백질 상호작용과 관련된 단백질 서열정보를 데이터베이스화하여 단백질 상호작용 기반의 단백질 기능분석에 널리 이용되는 대표적인 데이터베이스들로는 DIP (Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/)[9], BIND (the Biomolecular Interaction Network Database, http://www.blueprint.org/bind/bind.php)[2], GRID (the General Repository for Interaction Datasets, http://biodata.mshri.on.ca/grid/servlet/Index)[4] 등이 있다. DIP은 단백질 상호작용 데이터베이스 중에서 가장 널리 알려진 것으로 현재 15,114개의 단백질 쌍 레코드가 저장되어 있다. BIND는 상호작용하는 단백질 쌍 뿐만 아니라, 분자 복합체 및 대사경로에 관한 정보들도 저장하고 있는데, 현재 11,237개의 단백질 쌍 레코드를 갖고 있다. 마지막으로 GRID는 기존에 밝혀진 상호작용 단백질 쌍 데이터들을 통합하기 위해서 만든 데이터베이스로 20,984개의 상호작용 단백질 쌍 레코드를 갖고 있으며, DIP와 BIND의 일부 데이터와 중복된다.

이 논문에서는 단백질의 상호작용 관계를 단백질 서열의 상동성 수준, 단백질의 도메인 수준, 단백질의 상호작용 수준 등 다양한 측면에서 단백질 상호작용과 기능을 규명하는데 도움을 줄 수 있는 WASPIFA (Web-based Assistant System for Protein-Protein Interaction and Function Analysis)에 대해 소개하고자 한다. WASPIFA 시스템은 단백질 상호작용

데이터베이스인 DIP, BIND, 및 GRID 등을 단순히 통합화시켜 놓은 것이 아니라, 다양한 전처리(Pre-processing) 과정을 통한 데이터마이닝에 의해 도메인 수준 및 서열 상동성 수준에서 단백질 상호작용 및 기능분석을 행 할 수 있도록 구성되어 있다. 이러한 시스템은 비록 개발이 완료되었지만 올바른 관리와 주기적인 갱신이 안되면 시스템으로서의 생명력을 상실하게 되는데, 이러한 점을 감안하여 시스템 내에 관리 및 자동 갱신 기능까지 추가하여 전문적인 생물정보학자가 아니더라도 쉽게 시스템을 유지 및 관리 할 수 있게 구현하였다.

재료 및 방법

WASPIFA 시스템의 전체 구성

WASPIFA 시스템은 전 세계적으로 해당 연구자들이 가장 널리 사용하는 대표적인 단백질 상호작용 데이터베이스인 DIP, BIND 및 GRID 등을 통합하여 하나의 독립적인 통합 단백질 상호작용 데이터베이스를 로컬 사이트에 구축하였다 (Fig. 1의 X-Large DB 부분에 해당). 게다가, 이렇게 구축된 통합 단백질 상호작용 데이터베이스의 각 엔트리 단백질 서열에 대해 전처리 과정을 통해 서열 유사성 검색과 도메인 분석을 하여 얻어진 유사성 및 도메인 정보들을 WASPIFA 시스템 내에 로컬 데이터베이스화하였다(Fig. 1의 Function, Domain Info. 부분에 해당). 이러한 총체적인 정보를 바탕으로 단백질 상호작용 관계 및 기능을 유추하여 사용자에게 보고할 수 있도록 시스템을 구성하고 구현하였다(Fig. 1 참조). 보다 구체적으로 설명하자면 Fig. 1에서 볼 수 있듯이 기존의 데이터베이스들을 통합화하는 작업뿐만 아니라 여러 전처리 작업을 통해서 얻어진 데이터 및 정보들을 로컬 데이터베이스화하여 사용자에게 총체적인 분석결과를 제공할 수 있도록 구성하였다. 전처리 작업은 크게 3개의 범주로 나누어 볼 수 있다. 첫째, 단백질 상호작용 데이터베이스에 포함되어 있는 단백질 엔트리 서열들을 COG (Clusters of Orthologous Groups of proteins)[7] 및GO (Gene Ontology)[5] 데이터베이스를 대상으로 상동성 검색을 통해 단백질의 기능 및 주석(Annotation)정보들을 추출하여 내부적으로 FA (Function & Annotation) 데이터베이스를 구축하였다 (Fig. 1 참조). 상동성 검색시에는 대표적 상동성 검색 프로그램인 BLAST[1]를 사용하였고, 이때 사용한 E-value (Expect value)는 각각10^{-3} 및 10^{-5}으로 하였으며, 가장 유사성이 높은 것을 채택하였다. 따라서 사용자는 단백질 상호작용의 관계뿐만 아니라 입력 단백질의 기능정보 및 주석정보 등을 동시에 확인할 수 있다. 둘째, 단백질 상호작용 데이터베이스에 포함되어 있는 단백질 엔트리 서열들에 대해 모티프/도메인 통합 검색 프로그램인 InterProScan[10]을 사용하여 모티프/도메인 통합 데이터베이스인 InterPro[6]을 대상으로 모티프/도메인 검색을 하였다. 이러한 검색 결과를 체계적으로 정리하여 Domain DB를 구축하였다 (Fig. 1 참조). 따라서 사용자는 단백질 상호작용의 관계뿐만 아니라 분석하고자 하는 입력 단백질의 도메인 정보까지도 확인할 수 있어 사용자가 분석결과를 통해 보다 올바른 판단을 내릴 수 있을 것이다. 셋째, 앞의 전처리 과정에서 얻어진 Domain DB와 단백질 상호작용 통합

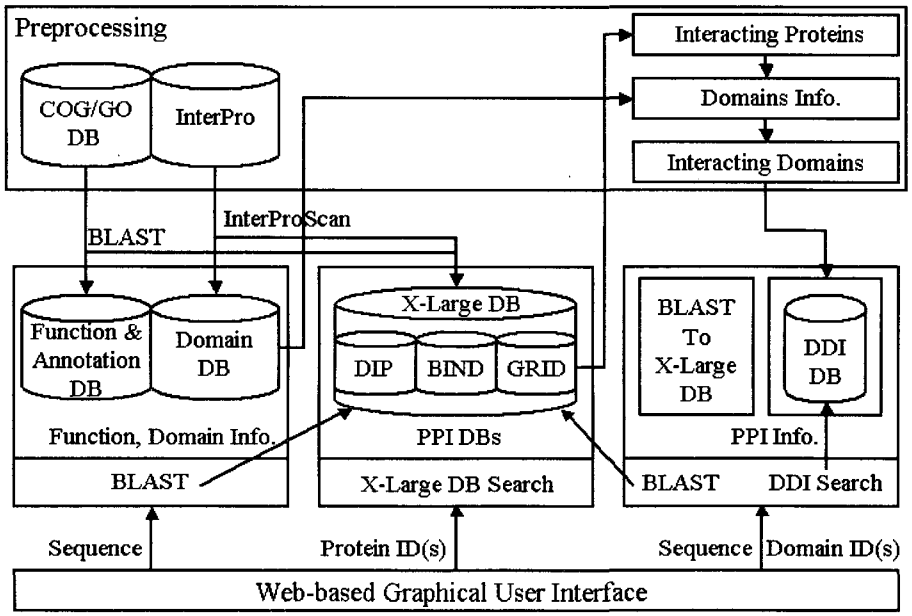

Fig. 1. Schematic diagram of WASPIFA system. The system consists of main three components - function/domain information, PPI (Protein - Protein Interaction) DBs, and PPI information, which are constructed by various preprocessing and using X-large DB. Arrows indicate work and information flows in the system.

데이터베이스내의 데이터를 이용해서 상호작용 단백질 쌍 정보와 그 단백질들이 갖고 있는 도메인 정보를 이용하여 DDI (Domain-Domain Interaction) 데이터베이스를 구축하였다(Fig. 1 참조). 다시 말해서, 두 단백질 사이의 상호작용은 그 두 단백질이 포함하고 있는 도메인 사이에서 일어난다는 것을 기반으로 상호작용하는 단백질과 그 단백질이 포함하는 도메인 정보를 이용하여 도메인 사이의 상호작용 쌍을 추출하여 데이터베이스화하였다. WASPIFA 시스템은 웹 기반으로 구현되었으며, 클라이언트, 서버 및 후미(back-end) 데이터베이스 등으로 구성되는 3-계층구조(3-tier architecture)를 갖고 있다. 서버쪽은 BLAST 서버 모듈이 핵심을 이루고, 후미 데이터베이스 부분은 앞에서 언급한 단백질 상호작용 통합 데이터베이스와 다양한 전처리 작업을 통해서 구축된 가공의 FA, Domain, 및 DDI 데이터베이스들로 구성되며, 클라이언트는 웹 기반의 GUI (Graphical User Interface)로 이루어져 있다. 본 시스템에서 사용한 데이터베이스 엔진은 MySQL (http://www.mysql.com)이다.

데이터베이스 통합화 및 데이터베이스 스키마

앞에서 언급한 것처럼 전 세계적으로 가장 대표적이며 많은 연구자들에 의해 널리 사용되는 단백질 상호작용 데이터베이스인 DIP, BIND 및 GRID 등을 통합화하여 자체적으로 하나의 단백질 상호작용 통합 데이터베이스인 X-Large 데이터베이스를 구축하였다. 이 논문에서는 데이터 구조가 서로 다른 데이터들을 통합하기 위해서 X-Large 테이블 (X_Large_T)을 생성하였다. X-large 테이블은 inter_id, pro1_id, pro2_id 및 db_type 필드 등으로 이루어진다. Inter_id는 X-Large 테이블에서 사용하는 단백질 상호작용 쌍의 식별자이고, pro1_id와 pro2_id는 상호작용하는 두 단백질의 식별자를 나타낸다. 이러한 식별자는 DIP 에서는 DIP 데이터베이스에서 사용되는 노드 아이디, BIND에서는 GI, GRID에서는 ORF 아이디이며 db_type은 단백질 상호작용 데이터베이스의 종류(DIP, BIND, GRID)를 표현한다. X-Large 테이블의 pro1_id와 pro2_id는 각 단백질들의 상세정보를 저장하고 있는 DIP_NODE_T, BIND_NODE_T, GRID_NODE_T의 단백질 아이디를 참조한다(Fig. 2 참조). 사용자는 X-Large DB를 이용해서 손쉽게 여러 개의 단백질 상호작용 데이터베이스를 검색할 수 있다.

그 밖에 Function과 Annotation, Domain, 및 Domain-Domain Interaction DB 스키마의 경우는 Fig. 3에 나타나 있는데, 기능정보 및 주석정보는 COG_T테이블과 GO_T 테이블의 COG 및 GO 식별자를 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 상호작용 하는 도메인 쌍과 해당 도메인을 갖는 단백질 정보를 함께 저장하고 있다(Fig. 3 참조).

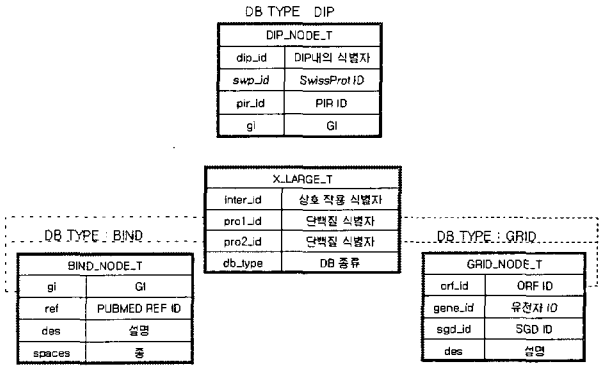

Fig. 2. Schema of X-Large DB. Dotted lines represent entity relationship between tables referenced by primary and foreign keys.

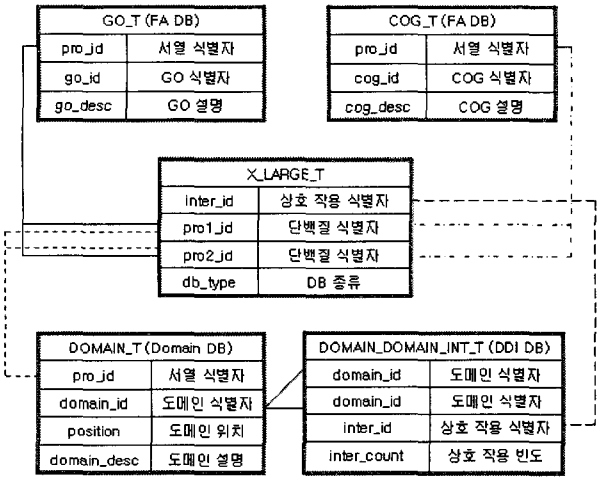

Fig. 3. Schema of FA (Function & Annotation), Domain, and DDI (Domain-Domain Interaction) databases. Normal and dotted lines represent entity relationship between tables referenced by primary and foreign keys.

X-Large DB 검색 기능

X-Large DB내의 DIP 데이터베이스에 대해서는 node 고유 번호 혹은 SWP ID, GI, PIR ID로 검색할 수 있게 구성하였고, 검색한 결과에는 검색한 단백질의 서열, COG로 나타나는 기능정보, GO로 표현되는 주석정보, 그리고 그 단백질이 포함하고 있는 도메인 정보들이 포함된다(Fig. 4). 물론, 검색한 단백질과 상호작용하는 단백질들을 확인할 수 있게 구현하였다. BIND 데이터베이스의 경우 사용자는 GI, Ref ID로 BIND에 포함되어 있는 단백질을 검색할 수 있다. 검색한 결과는 검색한 단백질의 종, 서열, COG ID, GO 그리고 포함한 도메인 정보들을 확인할 수 있다. 그리고 DIP의 경우처럼 검색한 단백질과 상호작용하는 단백질들을 확인할 수 있게 구성하였다. GRID 데이터베이스의 경우 사용자는 ORF ID, Gene ID, SGD ID로 검색할 수 있게 구현하였다. 검색 결과에는 DIP 검색에서와 마찬가지로 단백질의 서열, COG로

나타나는 기능정보, GO로 표현되는 주석정보, 그리고 그 단백질이 포함하고 있는 도메인 정보들이 포함된다. 또한 검색한 단백질과 상호작용하는 단백질들을 확인할 수 있다.

유사성 기반의 단백질 상호작용 및 기능 추론

유사성 기반은 사용자의 입력 단백질 서열을 X-Large 데이터베이스의 엔터리 서열들과 BLAST 상동성 검색을 하여 높은 유사성을 갖는 엔터리 서열을 찾는다. 이러한 단백질 엔터리 서열의 상호작용 관계를 기반으로 입력서열과 상호작용할 것으로 추정되는 단백질들을 검색할 수 있도록 시스템을 구현하였다. 이 방법은 단백질의 상호작용 관계를 전체 서열의 상동성에 의존하므로 상호작용에 실제로 관여하는 도메인 수준에서의 검토가 없다는 단점이 있지만, 일반적으로 주석전이(Annotation transfer) 측면에서 가장 일반적으로 사용되는 방법이라 할 수 있다. 사용자가 X-Large DB내의 해당 데이터베이스(즉, DIP, BIND, GRID 등)를 선택하고, BLAST의 E-value를 설정한 후 단백질 상호작용 관계 및 기능을 알고자 하는 입력 서열을 입력하면, 그 서열과 높은 유사성을 보이는 서열들을 데이터베이스별로 보여주고, 각 단백질의 ID를 클릭하면 해당 단백질의 상세정보와 그 단백질과 상호작용하는 단백질들을 확인할 수 있게 시스템이 구성되어 있다.

도메인 기반의 단백질 상호작용 및 기능 추론

앞서 언급했던 InterProScan과 InterPro를 이용하여 전처리 과정을 통해 구축한 DDI 데이터베이스를 기반으로 도메인 기반의 단백질 상호작용 및 기능을 분석할 수 있도록 WASPIFA 시스템을 구현하였다. 도메인 기반의 추론 방법에는 두 가지가 있다. 첫째는 사용자가 하나의 입력 도메인을 입력하는 것이고, 둘째는 사용자가 두 개의 도메인을 입력하는 것이다. 한 개의 도메인을 입력한 경우, 입력 도메인과 상호작용하는 DDI 데이터베이스의 관련 도메인 정보들을 보여주고, 두 개의 도메인을 입력한 경우에는 두 개의 도메인이 상호작용하는 빈도와 가능성 등을 보여주도록 구성하였다. 도메인 기반으로 상호작용하는지 여부의 척도가 되는 것은 자체 구축한 DDI 데이터베이스의 통계처리 값을 기반으로 한다.

관리 및 자동 갱신 기능

한 가지 주목해야 할 사항은 단백질 상호작용에 관한 데이터는 기하급수적으로 늘어나고 있기 때문에 전문 개발자가 아닌 일반 사용자가 시스템을 관리하고 최신 데이터를 갱신하는 것은 단순한 일이 아니다. 따라서 방대하게 증가하는 데이터들을 손쉽게 관리하고 갱신하기 위한 자동화 시스템을 설계하고 구현하여 WASPIFA 시스템에 추가하였다. 자동화 시스템의 개괄적인 흐름은 Fig. 5와 같다. 데이터를 업로

1. Node

Node ID	1000N	SWP ID	P12689
PIR ID	S67255	GI ID	2133088

2. COG

COG ID	KOG2093
COG Desp	Translesion DNA polymerase - REV1 deoxycytidyl transferase

3. GO

Num.	Name	Type	Acc.
1	nucleotidyltransferase	molecular_function	GO:0016779
2	DNA repair protein	molecular_function	GO:0003685
3	response to DNA damage	biological_process	GO:0006974
4	DNA repair	biological_process	GO:0006281
5	extrachromosomal circular DNA	cellular_component	GO:0005727

4. Domains

Num.	DB	ID	Name	Start	End	InterPro ID
1	3	PF00533	BRCT	163	249	IPR001357
2	3	PF00817	IMS	361	737	IPR001126
3	4	SM00292	BRCT	163	239	IPR001357
4	5	PS50172	BRCT	161	249	IPR001357
5	5	PS50173	UMUC	358	554	IPR001126

5. Interacting Proteins

Edge	Pro ID	Pro Name	Pro Name	Function	GO
857E	663N	SNP1	YIL061C	KOG0113	GO:0016779 GO:0003685 GO:0006974 GO:0006281 GO:0005727

Fig. 4. Result screen of X-Large DB search. User-friendly web interface was implemented to facilitate the end-user to understand easily search result. End-user can click the arrow button to retrieve more detail information on each item.

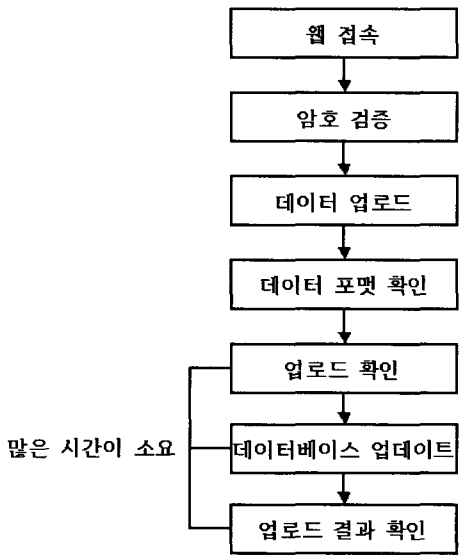

Fig. 5. Overall procedure for data update in WASPIFA system. The update procedure consists of seven steps including authentication one, which are indicated by normal line boxes.

드 하기 위해서 사용자는 우선 웹을 통해 WASPIFA 시스템에 접속을 하고 암호검증을 통해 데이터 업로드를 할 수 있

는 권한을 부여 받아야 한다. 데이터 업로드는 시스템 부하가 많이 걸리고 보안의 문제가 있을 수 있으므로 일반 사용자에게는 허용을 하지 않고, 관리자 권한을 가진 사용자만이 데이터를 갱신할 수 있도록 하였다. 권한이 인정된 사용자는 업로드하고자 하는 데이터의 형식에 맞게 업로드를 하고 WASPIFA 시스템은 데이터 형식이 갱신 가능한 형태인지를 확인한 후 사용자에게 업로드 확인과정을 보여준다. 사용자가 업로드 확인에 동의하면 WASPIFA 시스템은 데이터에 적절한 과정을 거쳐서 데이터베이스를 갱신하게 된다. 단순히, 단백질 상호작용 관련 데이터의 경우는 데이터 갱신이 바로 이루어질 수 있지만, FA (Function and Annotation) DB와 Domains DB의 경우는 유사성 검색 혹은 InterProScan을 이용해야 하므로 많은 시간이 소요된다 (짧게는 2~3일 길게는 1주일 이상 소용됨). 따라서 데이터 갱신의 결과는 바로 확인하기는 어려우며 시스템에서 제시하는 시간이 흐른 후에 확인이 가능하다. 사용자가 데이터를 업로드하면 WASPIFA 시스템은 내부에서 Fig. 6와 같은 과정을 거쳐서 데이터를 갱신하게 된다. 우선 단백질 상호작용 데이터 즉, 단순히 두 단백질 사이의 상호작용 관계를 표현하는 데이터는 데이터의 파싱(Parsing) 과정을 거치면 바로 데이터베이스에 갱신이 가능하도록 설계하고 구현하였다. Function DB와 Annoation DB의 경우는 서열 데이터를 받아서 COG와 GO 데이터베이스와 서열 유사성 검색을 하고 그 결과 COG ID와 Term ID를 얻어내고 그것을 데이터베이스화 하도록 구성되어 있다. Domains DB는 입력 받은 서열데이터에 대한 도메인 정보를 InterPro 데이터베이스를 대상으로 InterProScan 프로그램을 구동시켜 얻어낸 것인데, 이 과정의 작업은 아주 많은 시간이 소요된다. 그리고 DDI DB는 단백질 상호작용 관계정보와 도메인 정보를 이용해서 도메인 사이의 상호작용 관계를 추출하고 그 정보를 데이터베이스화 하도록 구현하였다.

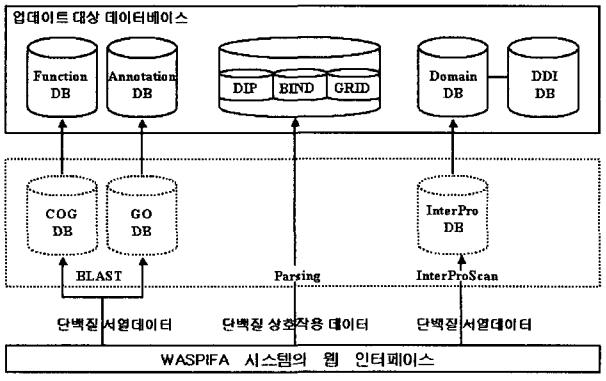

Fig. 6. Schematic diagram of overall procedure for updating databases involved. Normal line cylinders represent secondary databases, which are constructed from primary databases indicated by dotted line cylinders. Arrows mean data flow and preprocessing.

결론 및 고찰

본 논문은 단백질의 기능분석을 위해 핵심적으로 요구되는 단백질 상호작용 관계 및 기능정보 등을 체계적으로 제공할 수 있는 WASPIFA 시스템에 대해서 다루었다. WASPIFA 시스템은 기존의 단백질 기능 및 상호작용 관련 시스템과는 달리 분석하고자 하는 서열의 종합적인 정보 즉, 기능정보 및 주석정보, 도메인 정보, 상호작용 관계정보를 한 눈에 보기 쉽게 제공하기 때문에, 분석 대상서열의 종합적인 정보를 얻고자 하는 사용자가 편리하게 이용할 수 있는 시스템이라 할 수 있다. WASPIFA 시스템의 기능과 특징을 보다 더 구체적으로 언급하면 다음과 같이 요약할 수 있다. 기존의 단백질 상호작용 데이터베이스들을 통합하여 사용자가 한번에 손쉽게 단백질 상호작용 데이터를 검색 및 분석할 수 있다. 그리고 단백질 상호작용 데이터베이스의 ID나 서열의 상동성 검색을 통해 단백질의 상호작용 정보뿐만 아니라, 부가적으로 단백질의 기능정보 및 주석정보, 그리고 단백질이 포함하고 있는 도메인 및 모티프 정보들을 확인할 수 있으므로 단백질 서열을 종합적으로 분석할 수 있다. 게다가, WASPIFA 시스템은 전처리 과정에서 단백질 수준에서의 상호작용 관계정보를 이용해서 실제로 상호작용이 일어나는 도메인 수준에서의 상호작용 관계정보를 데이터베이스화하여 사용자는 손쉽게 도메인 수준에서의 상호작용 정보를 추출할 수 있다. 특히 중요한 특징 중의 하나는 빠른 시간 내에 기하급수적으로 증가하는 데이터를 효율적으로 저장, 갱신할 수 있는 내부 시스템을 갖추고 있다는 것이다. 신규 데이터의 갱신은 웹을 통해 아주 편리하게 이루어질 수 있는데, 현재 많은 데이터베이스들이 나날이 증가하는 단백질 상호작용 관련 데이터들을 갱신하여 제공하고 있지만 WASPIFA처럼 전처리 과정을 통해 단백질 상호작용에 관련된 부가적인 정보들 즉, 기능정보 및 주석정보, 도메인 정보, 도메인 사이의 상호작용 관계정보 등을 갱신하여 제공하지는 않는다. 그러나 우리가 개발한 시스템은 기존의 공개된 신규 데이터베이스의 데이터들뿐만 아니라 다양한 전처리 과정에 의한 데이터마이닝으로 얻어진 부가적인 데이터 및 정보들에 대해서도 자동 갱신될 수 있도록 설계 및 구현되어 쉽게 관리 될 수 있는 장점을 갖는다.

앞에서 언급한 것처럼 WASPIFA 시스템은 단백질의 기능분석을 위해서 단백질의 종합적인 정보를 얻고자 하는 사용자에게 큰 도움이 될 것으로 보이며, 앞으로는 도메인 수준에서의 상호작용 관계정보를 바탕으로 보다 정확하고 신뢰할 만한 상호작용 도메인 쌍을 밝혀내야 할 것이다. 현재 본 시스템은 공공 서비스 및 사용자의 편의성을 고려하여 웹 기반으로 개발되어 있어 일반 연구자들이 편리하게 이용할 수 있으나, 향후에 다양한 그래픽 라이브러리들을 사용하여 분석 결과를 보다 더 시각화 및 역동화 시켜야 할 것으로 여겨

진다. 그리고 상호작용 관계를 추론함에 있어서 그 가능성을 좀 더 세밀하고 정확하게 측정할 수 있는 방법이 도입되어야 할 것이다.

요 약

이 논문은 단백질의 기능분석을 위해 핵심적으로 요구되는 단백질 상호작용 관계정보 및 기능정보 등을 체계적으로 제공할 수 있는 WASPIFA (Web-based Assistant System for Protein-protein Interaction and Function Analysis) 시스템에 대해서 다루고 있다. WASPIFA 시스템은 특정 분야에 국한해서 단편적 정보를 제공하는 기존의 단백질 기능 및 상호작용 분석 시스템과는 달리 분석하고자 하는 서열의 종합적인 정보 즉, 기능정보 및 주석정보, 도메인 정보, 상호작용 관계정보 등을 제공한다. 일반 검색 및 분석 시스템에서 제공하지 못하는 종합적인 정보들은 다양한 전처리 과정을 통해서 얻어진 데이터 및 정보 등을 시스템 내에 로컬 데이터베이스화해 놓은 것이다. 최종 사용자는 종합적인 정보를 통해서 올바른 평가와 판단을 통해서 효과적인 단백질 상호작용 분석과 기능분석을 행할 수 있다. 또한 자동관리 및 데이터 갱신 기능을 갖추고 있어 시스템 관리자가 효율적으로 시스템을 유지 및 관리할 수 있다.

참 고 문 헌

1. Altschul, S. F., W. Gish, W. Miller and E. W. Myers. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
2. Bader, G. D., D. Betel and C. W. Hogue. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248-250.
3. Bock, J. R. and A. D. Gough. 2001. Predicting protein-protein interactions from primary structure. *Bioinformatics* **17**, 455-460.
4. Breitkreutz, B. J., C. Stark and M. Tyers. 2003. The GRID : The General Repository for Interaction Datasets. *Genome Biology* 4, R23.
5. Hennig, S., D. Groth, and H. Lehrach. 2003. Automated gene ontology annotation for anonymous sequence data. *Nucleic Acids Res.* **31**, 3712-3715.
6. Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. A. Sigrist, R. Vaughan and E. M. Zdobnov. 2003. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315-318.
7. Tatusov, R. L., D. A. Natale, I. V. Garkavtsey, T. A. Tatusoya, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova and E. V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22-28.
8. Valencia, A. and F. Pazos. 2002. Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology* **12**, 368-373.
9. Xenarios, I., L. Salwinski, J. Duan, P. Higney, S. Kim and D. Eisenberg. 2002. DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303-305.
10. Zdobnov, E. M. and R. Apweiler. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848.