Journal of Life Science 2008 Vol. 18. No. 3. 318~322 ©JLS / ISSN 1225-9918

단백질 상호작용 데이터의 효율적 관리와 자동 갱신을 위한 시스템 설계와 구현

김 기 봉*

상명대학교 공과대학 생명정보공학과

Received January 7, 2008 / Accepted March 20, 2008

System Design and Implementation for the Efficient Management and Automatic Update of Protein-Protein Interaction Data. Ki-Bong Kim*. *Department of Bioinformatics Engineering, Sangmyung University, Chunan 330-720, Korea* - This paper deals with an efficient management and automatic update sub-system for WASPIFA (Web-based Assistant System for Protein-protein Interaction and Function Analysis) system that had been developed in the past and now provides the comprehensive information on protein-protein interaction and protein function. Protein interacting data has increased exponentially, so that it costs enormous time and effort. In other words, it is actually impossible to manually update and manage an analysis system based on protein interacting data. Even though there exists a good analysis system, it could be useless if it was able to be updated timely and managed properly. Unfortunately, in most cases, biologists without professional knowledge on their analysis systems have to cope with a great difficulty in running them. In this respect, the efficient management and automatic update subsystem of protein interacting and its related data has been developed to facilitate experimental biologists as well as bioinformaticians to update and manage the WASPIFA system.

Key words : WASPIFA, protein-protein interaction, protein function, bioinformaticians, analysis system

서 론

유전암호 해독기술과 생명정보학 분야의 급속한 발전으로 엄청난 유전정보가 쏟아지고 있다. 2007년 11월 초에 IntegratedGenomics 사가 자신들의 웹 사이트(http://ergo.integratedgenomics.com/ERGO/)를 통해 밝힌 바에 의하면 1012여종의 완전 해독된 유전체 정보에 대한 서비스를 제공하고 있다고 한다. 이처럼 방대하고 다양한 유전체 서열이 밝혀짐에 따라 생명정보학 분야에서는 그러한 서열들이 암호화하고 있는 단백질들을 밝혀내고 해당 단백질들의 기능을 전산학적 기법으로 규명하고자 하는 연구가 엄청나게 이뤄지고 있다. 최근 얼마 전까지만 해도 단백질체(Proteome) 기능연구는 각 개별 단백질의 기능을 밝히는 것에 주안점을 두고 연구가 진행되어 왔다. 그러나 최근에는 개별 유기체의 전체 유전체 서열정보가 밝혀짐에 따라 단백질이 하나의 독립체로서 그 기능을 수행하는 측면뿐만 아니라, 전체적인 단백질 상호작용 네트워크의 구성요소로서 그 기능을 수행한다는 점에 초점을 맞추어 연구가 진행되고 있다. 게다가, 실용적인 측면에서는 특정 단백질간의 작용과 반작용은 신약 개발의 중요한 단서를 제공한다. 따라서 연구자들은 다양한 실험적인 방법과 전산학적인 방법들을 사용하여 단백질들 간의 상호작용 관계를 규명하려 한다[1,11]. 상호작용 단백질 쌍들에 대한 데이터와 부가정보 등을 체계적으로 모아 놓은 단백질 상호작용 데이터베이스로는 DIP (Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/) [8], BIND (the Biomolecular Interaction Network Database, http://www.bind.ca) [3], GRID (the General Repository for Interaction Datasets, http://www.thebiogrid.org/index.php) [9] 등이 대표적이다. DIP은 단백질 상호작용 데이터베이스 중에서 가장 널리 알려진 데이터베이스로써 56,186개의 단백질 상호작용 정보를 포함하고 있다[8]. BIND는 상호작용하는 단백질 쌍 뿐만 아니라, 분자복합체(Molecular complex)들과 대사경로(Pathway) 등에 관한 정보들도 저장하고 있는데, 80,000여 개의 상호작용 정보를 포함하고 있다[3]. 마지막으로 GRID는 기존에 밝혀진 상호작용 단백질 쌍 데이터들을 통합하기 위해서 만든 데이터베이스로 199,367개의 상호작용 원시 데이터를 갖고 있다[9].

본 논문에서는 이미 개발 완료하여 발표한 적이 있는 WASPIFA (Web-based Assistant System for Protein-protein Interaction and Function Analysis) 시스템[4]을 효율적으로 관리하고 운용할 수 있도록 시스템 내부의 백엔드 데이터베이스(Back-end database)들을 자동 갱신 및 관리하기 위한 운영관리 시스템 설계 및 구현에 대해 소개하고자 한다. WASPIFA는 단백질의 상호작용 관계를 단백질 서열의 상동성 수준, 단백질의 도메인 수준, 단백질의 상호작용 수준 등의 다양한 측면에서 단백질 상호작용과 기능을 규명하기 위한 시스템으로 내부적으로 DIP, BIND 및 GRID 등을 내포하고 있을 뿐만 아니라 다양한 처리과정에 의해 재생성된 여러

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

가공된 데이터베이스들이 백엔드 데이터베이스로 활용되고 있다[4]. 일반적으로 생명정보학 분야에서 다양한 용도와 목적에 맞추어 분석 시스템들이 개발되지만, 가장 중요하면서도 일반 사용자들에게 쉽게 간과되는 것은 개발된 분석 시스템을 제대로 운영하고 관리하는 문제이다. 하나의 분석 시스템도 생명체처럼 부단히 진화하지 않으면 그 생명력을 잃을 수밖에 없다. 즉, 시스템 내부의 자체 로컬 데이터베이스를 구성하는 원본 외부 데이터베이스의 갱신이 이루어졌을 때 그것에 발맞추어 내부 로컬 데이터베이스가 최신 번전으로 갱신되지 않으면 어느 누구도 분석 시스템을 더 이상 신뢰하지 않을 것이며, 그것으로 인해 더 이상 사용되지 않고 도태될 것이다. 즉, 뛰어난 기능을 갖춘 분석 시스템 개발뿐만 아니라 시스템 내부 로컬 데이터베이스의 지속적인 갱신과 관리 등이 해당 분석 시스템의 생명력을 좌우한다. 이러한 문제는 현실적으로 보면 개발자와 사용자 및 관리자가 다른 경우가 일반적이기 때문에 야기된다고 할 수 있다. 대부분의 경우 위탁 및 용역에 의해서 시스템이 개발되다 보니 생명정보학 분야의 시스템 개발 전문가에 의해 개발이 이루어지고, 시스템이 제공하는 기능 및 특성에 대해서는 해박한 지식을 갖고 있을지는 모르지만 시스템 내부의 세부 구성 및 데이터 구조 등에 대해서는 전혀 문외한 일반 생물학 실험연구자들이 사용하고 운영하는 경우가 태반이다. 이러한 대부분의 경우에는 매뉴얼이나 각종 문서 등에 의존해서 유지 및 관리해야하는 상황이나 그러한 매뉴얼과 문서 등을 비전문가가 이해한다는 것은 그리 쉬운 일은 아니다. 블랙박스(Black box)화 되어 있는 분석 시스템의 세부 내부 구조와 각종 파싱엔진(Parsing engine) 등에 대해 전혀 모르더라도 효과적으로 내부 로컬데이터베이스를 갱신하고 관리할 수 있는 도구가 있다면 시스템 운영 및 관리가 엄청나게 용이해 질 것이다. 이러한 측면에서 본 논문에서는 WASPIFA 시스템을 효과적으로 관리하고 운용하기 위한 방안으로 개발된 단백질 상호작용 및 기능 분석용 로컬 데이터베이스 자동 갱신 및 관리용 하부 시스템에 대해 소개하고자 한다.

재료 및 방법

WASPIFA 시스템의 전체 구성

WASPIFA 시스템은 이전에 언급한 바와 같이 궁극적으로는 단백질 상호작용 기반의 단백질 기능 예측 시스템이지만, 세부적으로는 단백질 서열의 상동성 수준, 단백질의 도메인 수준, 단백질의 상호작용 수준 등의 다양한 측면에서 단백질 상호작용과 기능을 규명하기 위한 분석 시스템이다[4]. 분석 시스템의 핵심은 분석모듈과 더불어 내부의 중심을 이루고 있는 다양한 구성요소 데이터베이스들이다. 전 세계적으로 해당 연구자들이 가장 널리 사용하는 대표적인 단백질 상호작용 데이터베이스인 DIP, BIND 및 GRID 등을 파싱하여 재구성이나 통합화 과정을 통해 생성된 독자적인 단일 단백질 상호작용 데이터베이스가 WASPIFA 시스템에 플러그인(Plug-in)되어 있다(Fig. 1의 X-Large DB 부분에 해당). 게다가, 이렇게 구축된 통합 단백질 상호작용 데이터베이스의 각 엔트리 단백질 서열에 대해 전처리(Pre-processing) 및 파싱 과정을 통해 서열 유사성 검색과 도메인 분석을 하여 얻어진 유사성 및 도메인 정보 등이 WASPIFA 시스템 내에 로컬 데이터베이스화되어 있다(Fig. 1의 Function DB, Annotation DB, Domain DB, DDI DB 등에 해당). 이러한 총체적인 정보를 바탕으로 단백질 상호작용 관계 및 기능을 유추하여 사용자에게 보고할 수 있도록 시스템이 이루어져 있다(Fig. 1). 즉, 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 구성되어 있다[4].

자동갱신 및 관리 대상 데이터베이스들

WASPIFA 시스템은 사용자에게 단백질의 상호작용 관계에 관한 유용한 정보를 제공하기 위해서 앞에서 언급한 것처럼 다양한 전처리 및 파싱 과정을 통해 크게 5개의 재가공된 데이터베이스들을 생성하여 내장하고 있다(Fig. 1). 시스템 내의 분석모듈이야 한 번 제대로 잘 개발되면 큰 문제가 없는 한 갱신하지 않아도 되지만 데이터베이스 부분은 부단히 갱신하고 관리해야 하는 부분이다. 이러한 대상 데이터베이스를 살펴보면 다음과 같다. 첫째, 기존의 단백질 상호 작용 관련 데이터베이스인 DIP, BIND 그리고 GRID를 통합한 X-Large DB이다. 둘째, 단백질의 상호작용 관계를 유추하기 위해서는 단백질의 기능 정보도 유용하다는 점에 착안하여 X-Large DB에 있는 단백질들을 COG (Clusters of Orthologous Groups of proteins) [10]를 대상으로 유사성 검색을 통해 각 단백질들의 기능 정보들을 추출하여 X-Large DB에 포함되어 있는 단백질들의 기능 정보를 모아놓은 Function DB이다. 셋째, Function DB와 마찬가지로 X-Large DB에 있는 단백질들의

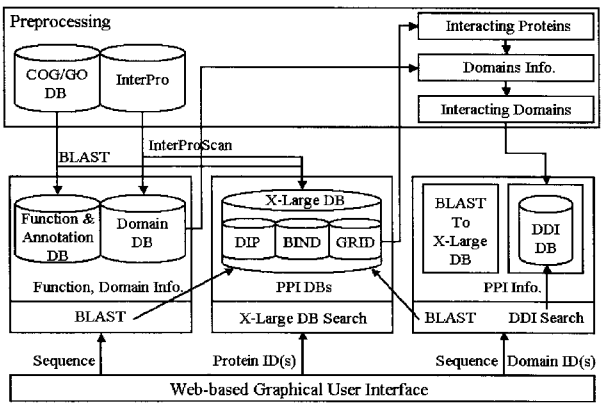

Fig. 1. Schematic diagram of WASPIFA system.

유전자 온톨로지(Gene ontology) 정보를 알 수 있도록 GO (Gene Ontology) [2]를 대상으로 유사성 검색을 하여 X-Large DB에 있는 단백질들의 주석정보 등을 모아 놓은 Annotation DB이다. 넷째, X-Large DB의 각 엔트리 서열들에 대해 단백질 도메인 데이터베이스인 InterPro [6]를 대상으로 도메인 탐색 프로그램인 InterProScan [7]을 실행시켜 얻은 각 엔트리들의 도메인 정보들을 추출하여 모아 놓은 Domains DB이다. 마지막으로 DDI (Domain-Domain Interaction) DB인데, 이것은 X-Large DB에서 단백질 상호작용 관계에 관한 정보와 Domains DB에 있는 도메인 정보를 이용하여 도메인 간의 상호작용 관계 정보를 추출하여 그것을 데이터베이스화 해놓은 것이다.

자동 갱신 시스템의 설계 및 구현

여기서 한 가지 주목해야 할 사항은 단백질 상호작용에 관한 데이터는 기하급수적으로 늘어나고 있는데 그러한 데이터들을 사용자가 수동으로 관리하고 갱신하는 작업은 현실적으로 거의 불가능하다는 것이다. 또한 앞에서 언급한 WASPIFA 시스템의 내부 구성을 보면 알겠지만 각각의 구성 데이터베이스들이 실타래처럼 서로 얽혀 있어서 하나가 바뀌면 마치 도미노 현상처럼 연쇄적으로 다 수정되어야 한다. 따라서 본 논문에서는 방대하게 증가하는 데이터들을 손쉽게 관리하고 갱신하기 위한 자동 갱신 및 관리 시스템을 설계하고 구현하였다. 자동 갱신 및 관리 시스템의 개괄적인 작업 진행 순서는 Fig. 2와 같다.

데이터를 갱신하고 업로드하기 위해서 사용자는 우선 웹상의 WASPIFA 시스템에 접속하고 암호 검증을 통해 사용자 인증을 받아야 한다. 데이터를 갱신하고 업로드하는 과정은 시스템 부하가 많이 걸리고 시스템 보안의 문제가 있을 수 있으므로 일반 사용자에게는 권한 허용을 하지 않고, 관리자 권한을 가진 사용자만이 데이터를 갱신하고 업로드 할 수 있도록 구성하였다. 권한이 인정된 사용자는 업로드 하고자 하는 데이터의 양식에 맞게 업로드를 하면 WASPIFA 시스템은 데이터 양식이 갱신 가능한 양식인지 확인한 후 사용자에게 업로드 확인 과정을 보여준다. 사용자가 업로드 확인에 동의하면 WASPIFA 시스템은 블랙박스화 되어 있는 내부의 여러 전처리 과정 및 가공과정을 거쳐서 해당 데이터베이스들을 갱신하도록 설계 및 구현되었다. X-Large DB의 경우 입력 양식에 맞게끔 최신의 DIP, BIND, 및 GRID를 다운로드 받아서 업로드 시키면 내부적으로 재구성 및 파싱 과정을 거치기 때문에 상대적으로 짧은 시간 내에 갱신 작업이 이뤄진다. 그러나, Function DB, Annotation DB, Domains DB, 및 DDI DB 등의 경우는 BLAST를 이용한 유사성 검색이나 InterProScan 프로그램을 이용한 도메인 검색을 하여야 함으로 오랜 시간이 소요될 수 밖에 없다(시스템의 성능과 입력 데이터의 양에 따라 다를 수 있겠지만 일반적으로 짧게는 2~3일 길게는 1주일 이상 소용됨). 따라서 데이터 갱신의 결과는 바로 확인하기는 어려우며 시스템에서 제시하는 시간이 흐른 후에 확인이 가능하도록 구현하였다.

사용자가 데이터를 업로드 하면 WASPIFA 시스템은 내부에서 Fig. 3과 같은 과정을 거쳐서 데이터를 갱신하게 된다. 우선 X-Large DB 즉, 단순히 두 단백질 사이의 상호작용 관계를 표현하는 데이터는 위에서 언급한 바대로 재구성과 파싱 과정을 거치면 바로 갱신이 되도록 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 구성하였다(Fig. 4). Domains DB는 입력 받은 서열데이터를 InterPro 데이터베이스에 InterProScan을 통해 도메인 정보들을 얻어내는 것인데, 이 작업은 아주 많은 시

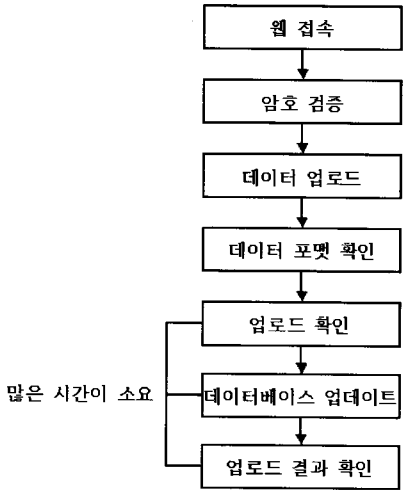

Fig. 2. Overall workflow of automaitc data update for WASPIFA system.

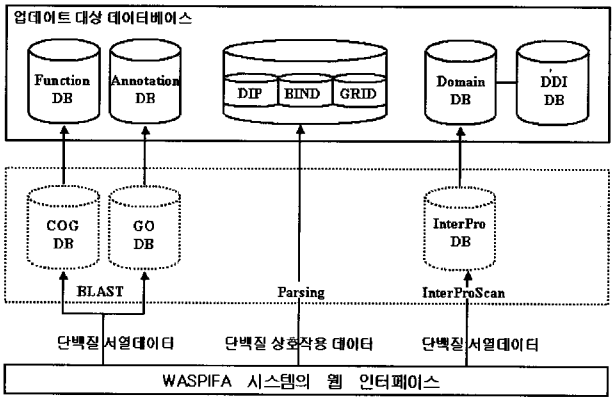

Fig. 3. Schematic diagram of overall procedure for updating databases involved. Normal line cylinders represent secondary databases, which are constructed from primary databases indicated by dotted line cylinders. Arrows mean data flow and preprocessing.

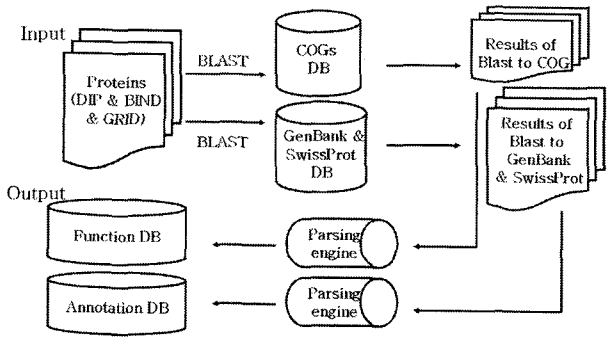

Fig. 4. Overall procedure for updating Function and Annotation database.

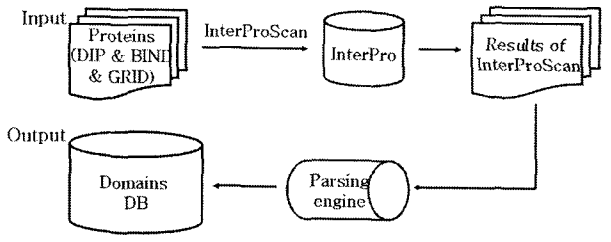

Fig. 5 Overall procedure for updating Domains database.

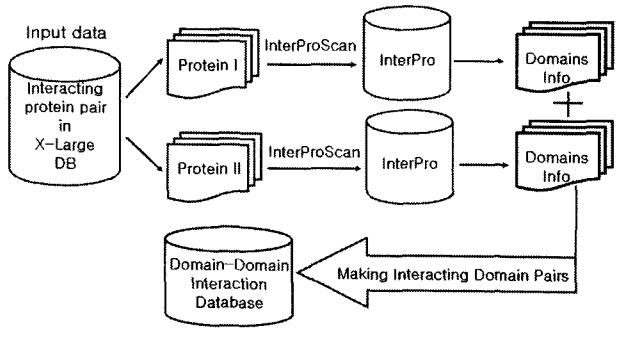

Fig. 6. Overall procedure for updating DDI (Domain-Domain Interactions) database.

간이 소요된다(Fig. 5). 그리고 DDI DB는 단백질 상호작용 관계 정보와 도메인 정보를 이용해서 도메인간의 상호작용 관계를 추출하고 그 정보를 데이터베이스화 하도록 구현하였다(Fig. 6).

결과 및 고찰

Fig. 7에서 알 수 있듯이 앞부분에서 언급된 복잡한 내부적 갱신 방법 및 절차 등을 모르더라도 누구나 쉽게 갱신 작업과 관리 작업 등을 수행할 수 있도록 웹 인터페이스를 구현하였다. 갱신하고자 하는 데이터 형식에 맞게 데이터를 업로드하고 데이터의 형식이나 업로드 방법에 관한 상세한 설명을 보고자 한다면 '설명' 버튼을 클릭해서 확인할 수 있다

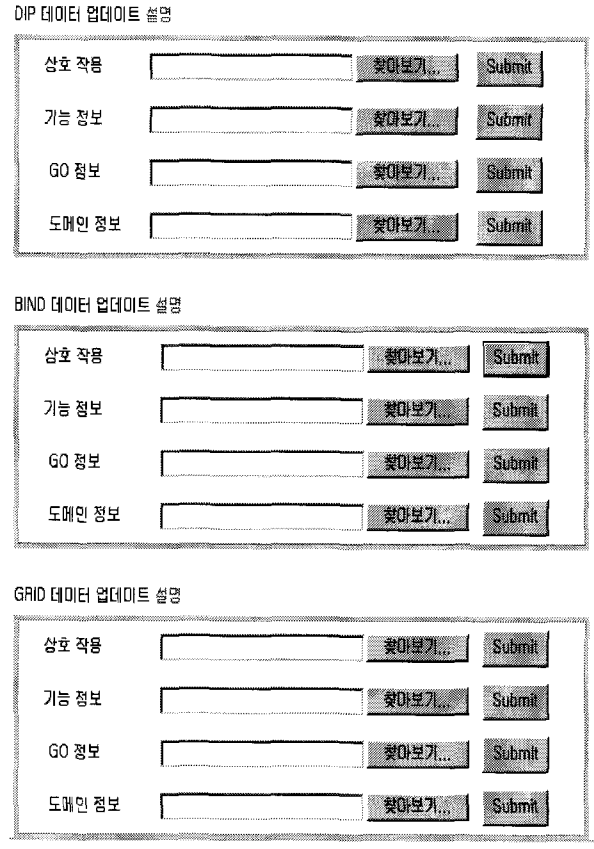

Fig. 7. Web-based interface for automatic data update and management.

(Fig. 7). 데이터가 업로드 되면 사용자에게 업로드 확인 과정을 거치고 데이터를 갱신하게 되는데, 데이터 갱신 시간이 많이 걸리는 경우는 일정 시간이 지난 후에 결과를 확인할 수 있다. 데이터 갱신 과정이 완료되면 Fig. 8에서 볼 수 있듯이 처리 결과를 확인할 수 있다. 업로드한 데이터의 수와 실제로 갱신이 이루어진 데이터의 수, 그리고 갱신이 제대로 이뤄지지 않은 데이터의 수 등을 일목요연하게 보여주도록 결과화면을 구성하고 구현하였다(Fig. 8).

DIP 상호 작용 데이터의 갱신 결과

Num	파일 이름	갱신 날짜	성공 수	실패 수
1	xx4.txt	07년07월13일10시11분	4	0
4	xx5.txt	07년09월07일09시55분	3222	0
5	xx3.txt	07년07월08일11시21분	12	0

DIP 서열 데이터의 갱신 결과

Num	파일 이름	갱신 날짜	성공 수	실패 수
2	dip1.seq	07년07월07일03시32분	2	0
3	yy4.txt	07년07월07일03시41분	4	0
4	yy5.txt	07년07월07일04시20분	4	0
5	yy.txt	07년07월14일03시10분	67	0

Fig. 8. User friendly web interface for reporting data update results.

본 논문에서는 단백질 상호작용 및 기능 분석 시스템인 WASPIFA를 효과적으로 운영 및 관리하기 위해 개발된 관련 데이터 자동 갱신 및 관리 하부시스템에 대해 소개하였다. WASPIFA 시스템을 이용하면 사용자는 단백질 상호작용 관련 데이터베이스를 통합 검색할 수 있으며 단백질의 상호작용 및 기능 분석을 유사성과 도메인 수준에서 유추해 볼 수 있다. 그러나 한 가지 고려해야 할 사항은 단백질 상호작용 관련 데이터들이 기하급수적으로 증가하고 있는데 그러한 데이터들을 수동으로 갱신하고 관리하는 작업은 엄청나게 많은 시간과 노력이 들뿐만 아니라 현실적으로 거의 불가능하다. 게다가 앞에서 언급한 바와 같이 대부분의 경우 분석 시스템을 도입하여 실제 사용하는 실험 생물학자들이 직접 시스템을 운영하고 관리해야 하는 경우가 대부분이기 때문에 관련 데이터의 갱신 및 관리 차원이 매우 중요하게 대두될 수 밖에 없다. 이러한 측면에서 값비싼 고가의 뛰어난 분석 시스템을 도입하고도 제대로 운영 및 관리가 이뤄지지 않아서 방치되는 경우가 현실적으로 비일비재하다. 이에 본 논문에서는 기하급수적으로 증가하는 데이터를 효율적으로 시스템에 새로이 반영하고 관리할 수 있는 하나의 방안을 WASPIFA 시스템을 대상으로 제시하였다. 데이터의 갱신은 웹을 통해 아주 편리하게 이루어질 수 있는데 현재 많은 데이터베이스들이 날로 증가하는 단백질 상호작용 관련 데이터들을 갱신하여 제공하고 있지만 WASPIFA 시스템처럼 전처리와 파싱 과정을 통해 단백질 상호작용에 관련된 부가적인 정보들 즉, 기능 및 주석 정보, 도메인 정보, 도메인 사이의 상호작용 관계 정보 등을 갱신하여 제공하지 않는다. 외부 원본 데이터의 변화에 효과적으로 대처하여 시스템에 잘 반영하고 축적한다면 WASPIFA 시스템의 분석 정확성도 훨씬 높아질 것이다. 즉, 데이터의 자동화된 갱신 및 관리 시스템은 사용자(특히 시스템 관리자)에게 매우 편리함을 줄 뿐만 아니라 분석 결과의 정확성도 높인다는 측면에서 파급효과가 매우 크다 할 것이다.

요 약

단백질 상호작용 관련 데이터들이 기하급수적으로 증가하고 있는데 그러한 데이터들을 수동으로 갱신하고 관리하는 작업은 엄청나게 많은 시간과 노력을 요구한다. 뿐만 아니라 개발자가 아닌 비전문가인 생물학자들이 시스템 구성 데이터베이스들을 갱신하고 관리하며 분석 시스템을 운영한다는 것은 현실적으로 거의 불가능하다. 이러한 측면에서 단백질 상호작용 정보를 이용한 효율적인 단백질 기능분석 시스템인 WASPIFA에 대해 자동적으로 데이터를 갱신하고 관리할 수 있는 시스템을 설계하고 개발하였다. WASPIFA 시스템은 단백질의 상호작용 관련 데이터들을 통합하여 사용자가 편리하게 데이터를 검색할 수 있으며 단백질 상호작용에 관련된 정보 즉, 기능 및 주석 정보, 도메인 정보, 도메인 간의 상호 작용 정보 등을 제공해 주는 유용한 단백질 기능분석 시스템이다.

References

1. Alfonso, V. and P. Florencio. 2002. Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology* **12,** 368-373.
2. Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258-D261.
3. Gilbert, D. 2005. Biomolecular interaction network database. *Briefings in bioinformatics* **6**, 194-198.
4. Jung, M., W. Park and K. B. Kim. 2004. Development of web-based assistant system for protein-protein interaction and function analysis. *Journal of Life Science* **14**, 997-1002.
5. Kim, K. B. 2007. A Study on the application methods of a support vector machine for gene promoter prediction. *Journal of Life Science* **17**, 714-718.
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. Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder and R. Apweiler. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116-W120.
8. Salwinski, L., C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg. 2004. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449-D451.
9. Stark, C., B. J. Breitkreutz, T. Requly, L. Boucher, A. Breitkreutz and M. Tyers. 2006. BioGRID: a general repository for integration datasets. *Nucleic Acids Res.* **34**, D535-D539.
10. Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 1-14.
11. Wan, K. and P. Jong. 2002. Large Scale statistical prediction of protein-protein interaction by Potentially Interacting Domain (PID) pair. *Genome Informatics* **13,** 42-50.