Journal of Life Science **2004**, Vol. 14. No. 6. 991~996 ©JLS

모티프 및 도메인 검색을 위한 통합 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 Integrated System for Motif and Domain Search. 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 an integrated system that facilitates researchers to do motif and domain search effectively and systematically. The system we developed is constructed on the basis of the integration of various resources related to motif, domain, and protein family. Those resources that can be classified into databases and search programs are dispersed to be available in Internet. In order to develop this system, we extracted core contents of diverse databases, which are required to analyze the protein function in terms of motifs or domains, to construct local databases and installed motif or domain search programs on our server, which corresponding database has as its own search program. Diverse utilities and CGI (Common Gateway Interface) programs make the databases and the search programs interlocked and web-based graphical user interfaces integrate all the components of our system. Employing our integrated system, end-users can receive its one-stop service to do protein function analysis systematically and effectively, without surfing many sites in Internet and wasting time over integrating search results.

Key words – Integrated system, motif, domain, local database, one-stop service, protein function analysis

서로 이질적인 분자생물학과 전산학을 근간으로 다양한 학문분야가 접목된 생물정보학은 고전적인 생물학적 기법으로부터 대용량 기법으로의 전환을 야기하고 있다. 생물정보학에 의한 올바른 유전자 기능분석 및 예측은 고전적인 생물학적 기법에 의한 신규 유전자 발굴 및 기능규명을 위해 요구되는 엄청난 비용을 절감하고, 또한 실험적으로 검증하기까지 소요되는 많은 시간을 획기적으로 단축시켜 줌으로써 지대한 경제적 파급효과를 줄 수 있다. 이러한 유전자 기능분석에는 유전자 구조 예측, 전사관련 신호부위 및 관련단백질인자 결합영역 예측, 프로모터 영역 예측, 단백질상의 특정 기능부위 예측, 기존의 핵산 및 단백질 서열 데이터와의 상동성 비교 등이 기본적으로 요구되며, 이러한 분석 결과들은 신규유전자 발굴 및 기능분석에 중요한 실마리를 제공한다[4,8,13]. 이 중에서도 진화론적으로 유관한 단백질이나 DNA 서열들이 기능 및 구조적으로 공유하는 공통의 특징에 대한 생물학적 실체인 모티프(Motif) 및 도메인(Domain)에 대한 검색은 유전체의 기능분석과 분류에 중요한 역할을 한다. 앞으로 이 분야의 지식기반이 확대됨에 따라 향후에는 그 비중이 더욱 더 커질 것으로 전망된다.

모티프 및 도메인 검색은 크게 두 가지 의미를 가진다. 첫째는 주어진 서열데이터 집합으로부터 그 집합을 대표할만한 서열이나 서열모델, 즉 추정되는 모티프 및 도메인을 찾아내는 것이다[2,12,15]. 정렬된 서열 데이터 집합을 입력으로 하거나, 정렬되지 않은 서열 데이터 집합을 입력으로 하여 그 집합을 대표할 수 있는 서열 및 서열모델을 만드는 다양한 알고리즘들이 세계적으로 이미 개발되어 사용되고 있다. 이러한 알고리즘들은 정규표현(Regular expression), 프로파일(Profile), 가중치 매트릭스(Weight matrix), 다중정렬(Multiple alignment), 정보이론(Information theory), EM(Expectation Maximization), 깁스 샘플링(Gibbs sampling), HMM (Hidden Markov Model) 등에 기반을 두고 있으며, 대표적인 실례가 MOTIF, MKDOM, MEME (Mutiple EM for Motif Elicitation), Meta-MEME 및 EMOTIF 등이 있다[2,7]. 이러한 알고리즘을 기반으로 모티프 및 도메인 관련 다양한 데이터베이스들이 구축되어 운용되고 있는 실정이다. 둘째는 주어진 서열이 특정 모티프나 도메인을 갖고 있는지 여부를 기존 공개용 데이터베이스를 대상으로 검색하여 관련된 데이터베이스 정보들을 찾는 것이다. 일반적으로 널리 사용되는 가장 대표적인 해당 데이터베이스로는 PROSITE[6], PRINTS[1], SMART[10], InterPro[11,14], Pfam[3], TIGRFam[9], 및 PRODOM[5] 등이 있다. 이들 데이터베이스는 각기 조금씩 다른 이론적인 배경이나 기준에 따라서 데이터를 분류하기 때문에 나름대로 개별적인 특색을 지니고 있다. 그러나 일반 연구자들이 검색결과를 직관적으로 이해하기에는 상당히 복잡하고 난해하다. 개별 데이터베이스의 특색을 최대한 반영하면서 동시에 통합적 검색이 가능한 서비스는 현재로서는

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

없는 실정이다. 또한 해당 데이터베이스와 연동되는 다양한 분석 도구들이 존재하나 상호간에는 유기적으로 연동되지 않는다. 다시 말하면, 개별 데이터베이스 별로 독립적인 사이트에서 독자적으로 운용되고 있기 때문에 사용자들이 통합적이고 총체적인 분석을 하기 위해서는 해당 사이트들을 일일이 찾아 옮겨 다니면서 검색하고, 검색 결과를 요약해야 하는 어려움이 있기 때문에 체계적이고 유기적인 검색을 위한 통합 시스템의 개발이 절실하다고 말할 수 있다.

이러한 측면에서 이 논문은 첫번째 사안이 아니라 두번째 사안에 대해 다루고 있다. 즉, 산재해 있는 활용 가능한 다양한 리소스들을 통합화하여 체계적이고 효율적인 모티프 및 도메인 검색을 할 수 있는 통합 시스템 개발에 대해 다루고자 한다. 이러한 시스템을 개발하기 위해 산재해 있는 모티프 및 도메인 관련 개별 데이터베이스의 핵심 부분만을 취합하여 로컬 데이터베이스화하고 단일 웹 인터페이스를 통해 통합하였고, 해당 데이터베이스에 딸린 개별 분석 도구들을 단일 웹 인터페이스 하에 통합하였으며, 단일 웹 기반 통합 인터페이스 상에서 검색 및 분석결과를 쉽게 이해할 수 있도록 일목요연하게 정리하여 사용자에게 보고하도록 시스템을 구현하였다. 국내 유전체 및 단백질 연구자들이 보다 효과적으로 활용할 수 있는 모티프 및 도메인 통합 검색 시스템을 구축함으로써 일반 연구자들이 인터넷 상에 산재해 있는 다양한 분석용 리소스들을 최대한 활용할 수 있어 개별 연구수행의 효율성을 극대화 시킬 수 있을 것이다.

재료 및 방법

통합 시스템의 전체구성

일반적으로 단백질은 모티프 및 도메인 등으로 모듈화 되어 있기 때문에 기능분석을 위해 기존의 데이터베이스에 들어있는 서열 데이터들과 상동성 검색을 할 경우 적중되지 않고 간과될 수 있다. 이러한 측면에서 위음성(false negative) 비율을 줄이기 위해서는 모티프 및 도메인 수준의 서열분석이 기능분석을 위한 필수적인 사안이 된다. 이러한 측면에서 우리가 개발한 모티프 및 도메인 검색을 위한 통합 시스템은 크게 세 개의 핵심요소로 이루어져 있다. 즉, 모티프 및 도메인 관련 데이터베이스 부분, 모티프 및 도메인 검색 프로그램 부분, 그리고 마지막으로 이러한 양자를 통합해 주고 내부적으로 상호참조에 의해 유기적으로 연동될 수 있도록 하는 인터페이스 및 각종 유틸리티 부분이다. 통합 시스템의 전체 구성을 보여주는 Fig. 1에서 알 수 있듯이 이 시스템은 웹 기반으로 이루어져 있으면 전체적으로는 3 계층 구조(3-tier architecture), 즉 클라이언트, 서버 및 백엔드 데이터베이스(Back-end database) 등으로 이루어져 있다. 특히, 데이터베이스 부분은 시스템 내에서 차지하는 비중이 상당히 크고 중요하기 때문에 보다 세분화하여 다룰 필요가 있다. 모티프 및 도메인 관련 데이터베이스들 중에서 가장 대표적인 InterPro 데이터베이스와 PROSITE 데이터베이스는 다른 데이터베이스들, 즉, PFam, ProDom, PRINTS, SMART, TIGRFam 등과는 다소 차별화하여 로컬 데이터베이스화하였다. 뿐만 아니라 이러한 데이터베이스들의 자료가 외부의 공개용 핵산/단백질 데이터베이스 및 MEDLINE 등과 하이퍼링크되게 시스템이 개발되었다. 서론에서 언급한 대로 모티프 및 도메인 관련 개별 데이터베이스들은 각각의 독립적인 사이트에서 운용되며, 개별 데이터베이스는 자신의 자료를 기반으로 모티프 및 도메인을 검색할 수 있는 프로그램을 갖고 있다. 본 시스템에서는 이러한 개별 프로그램들을 자체적으

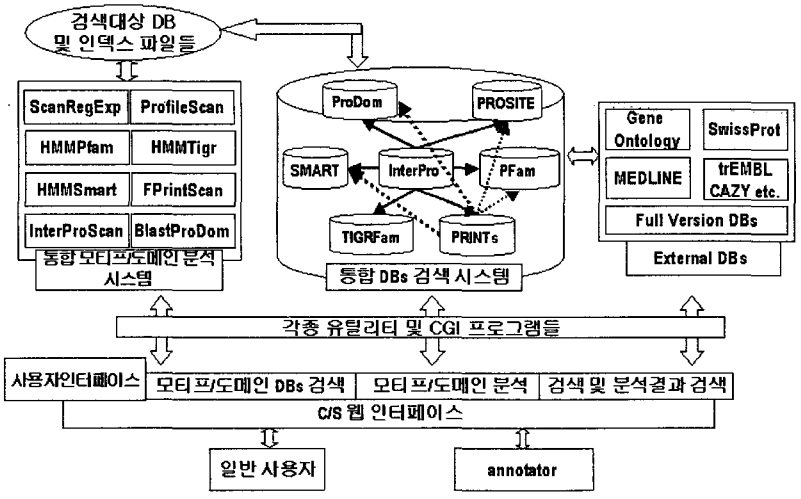

Fig. 1. Framework of the integrated system for motif and domain search. It consists of three main components, that is, databases (central cylinder and right rectangle parts), search programs (left rectangle part), and integrated user interface and diverse utilities. Various utilities including CGI programs interlock databases, search programs and user interfaces. Each thick arrow represents the relationship among sub-modules. The arrows in the integrated databases mean cross-references between databases.

로 인스톨하고 해당 CGI (Common Gateway Interface) 프로그램을 통해 웹 인터페이스 상으로 분석 서비스를 제공할 수 있도록 시스템을 구성하였다. Fig. 1에서 볼 수 있듯이 통합 모티프/도메인 분석시스템 부분에 해당되는 프로그램들, 즉, ScanRegExp, ProfileScan, HMMPfam, HMMTigr, HMMSmart, FprintScan, InterProScan, 및 BlastProdom 등이 이에 해당한다.

모티프 및 도메인 관련 데이터베이스들

앞에서 언급한 것처럼 본 연구의 통합 검색 시스템에서 데이터베이스 부분은 상당히 큰 비중을 차지하며, 실제로 매우 중요한 역할을 한다. Fig. 1에서 "통합 DBs 검색 시스템" 부분과 "External DBs" 부분이 시스템 내의 데이터베이스 영역에 해당한다. "External DBs" 부분은 자체적으로 로컬 사이트에 구축된 것이 아니고 하이퍼링크를 통해 시스템과 내부적으로 연결된다. "통합 DBs 검색 시스템" 부분에 속하는 데이터베이스들은 모두가 모티프 및 도메인 관련 데이터베이스로서, 이들은 일반 유전체 및 단백질 연구자들이 유전자 산물, 즉 단백질의 기능분석을 모티프 및 도메인 기반으로 행할 경우 일반적으로 널리 사용되는 것들이다. 이러한 측면에서 이들을 로컬 데이터베이스화하였고, 개별 데이터베이스에 대한 주요 필드별 검색 기능을 구현하였다. 이들 데이터베이스의 내용은 상당히 난해하고 복잡하기 때문에 단백질 기능분석에 필수적인 부분만을 추출하고 파싱(parsing)하여 축약된 형태의 데이터베이스로 구축하였고(이 논문에서는 앞으로 축약 데이터베이스라 칭함), 이에 대해 사용자가 핵심적인 주요 필드에 대해 일차적인 검색을 할 수 있도록 하였다. 핵심적이고 요약적인 정보가 제공되는 일차적인 검색결과는 외부의 풀버전(full version) 데이터베이스와 하이퍼링크 되게 구현하여, 사용자가 원할 경우에는 언제든지 검색결과의 해당 엔트리에 대한 전체 내용에 손쉽게 접근할 수 있게 시스템을 구축하였다. 본 시스템에서 축약된 형태의 데이터베이스로 구축된 것은 SMART, PRINTS, ProDom, TIGRFams, 및 PFam 등이고, 이들에 대해서는 Accession Number와 Text (Description Field 부분에 해당)에 대해 일차적인 검색이 가능하다. 모티프 및 도메인 관련 데이터베이스들 중에서 핵심적이고 대표적인 데이터베이스에 속하는 PROSITE의 경우 비록 축약된 형태이긴 하나 PDOC을 통해서 원래의 외부 "Document" 부분과 하이퍼링크 되게 구현하였다. 모티프, 도메인, 및 단백질 패밀리 관련 데이터베이스들을 통합화한 InterPro 데이터베이스의 경우 풀버전을 다운로드 받아 로컬 데이터베이스화하였다. 이에 대한 내용은 뒤에 나오는 "InterPro 데이터베이스 및 InterProScan" 영역에서 자세히 언급하고 있다. 본 시스템에서는 MySQL을 DBMS (Database Management System)로 사용하였고, 원시 데이터나 검색결과를 파싱하거나 재포맷(Reformat)하는 유틸리티나 각종 CGI 프로그램들은 Perl, C 언어 및 쉘 스크립트(Shell script) 등으로 구현되었다. 축약 데이터베이스인 SMART, ProDom, PFam, TIGRFam, Prosite 및 PRINTS 등은 각각 하나의 테이블로 할당하여 전체적인 데이터베이스 스키마(schema)를 구성하였다(Fig. 2). Fig. 2에서 볼 수 있듯이 각 테이블의 필드 구성은 단백질 기능분석을 위해서 필수적으로 요구되는 것들로 구성하였다. 이러한 축약 데이터베이스들은 InteProScan[5] 프로그램 배포판에 내장된 검색 대상 데이터베이스 부분을 정밀하게 분석하고, 복잡한 파싱(parsing) 및 재포맷(reformatting) 과정을 거쳐서 해당 필드 부분을 뽑아내어 데이터베이스화하였다. 해당 테이블의 Primary Key값을 통해서 원래 운용 사이트의 풀버전 레코드와 하이퍼링크된다.

데이터베이스 검색 기능

앞 부분에서 언급한 모티프 및 도메인 관련 축약 데이터베이스에 대한 검색 기능이란 두 가지 의미를 갖고 있다. 하나는 데이터베이스에 대한 필드 검색을 의미하는 것과 또 다른 하나는 입력 단백질 서열이 주어졌을 경우 해당 데이터베이스를 이용하여 입력 서열상에 모티프 및 도메인이 포함되어 있는지 여부를 검색하는 경우이다. 첫 번째의 경우에 있어서는 앞에서 언급한 것처럼 필드 검색, 즉 primary key값에 해당되는 accession number에 대한 검색과 Text 검색이 가능하도록 구현하였다. 앞에서 언급했듯이 검색된 해당 레코드가 웹 인터페이스 상에 제공되고, 하이퍼링크를 통해서 원래 사이트의 풀버전 레코드에 접근할 수 있도록 구현하였다. 이렇게 함으로써 일차적인 검색 정보 필터 기능을 사용자에게 제공한다. 그래서 사용자가 쓸모없는 정보의 홍수 속에서 혼돈이 없도록 핵심 정보만 제공하고, 그것이 원하는

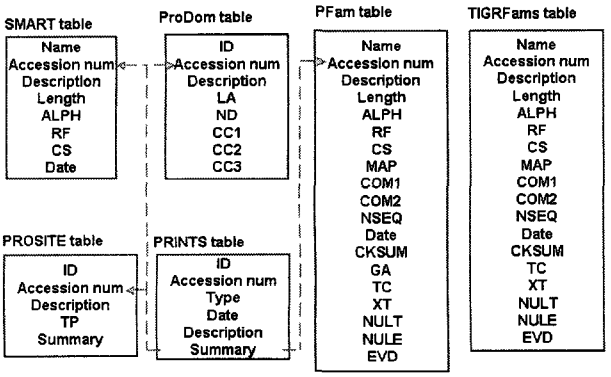

Fig. 2. Schema of abbreviated database. It consists of 6 tables that correspond to SMART, PROSITE, ProDom, PRINTS, Pfam, and TIGRFams, respectively. Each summarized record in this abbreviated database is hyper-linked with original record in external full version database by means of each primary key in individual table. The dotted arrows represent the relationship between tables referenced by primary key.

정보이면 보다 광범위한 정보에 용이하게 접근할 수 있게 설계하고 구성하였다. 두 번째 경우의 것은 각 축약 데이터베이스들이 각각의 해당 검색 프로그램들을 갖고 있는데, 이러한 검색 프로그램들의 내부 알고리즘을 일반 사용자들이 이해한다는 것은 매우 어려운 일이다. 그래서 개발자 입장에서 최대한 사용자들이 용이하게 사용할 수 있도록 인터페이스를 강화하고 소스코드(source code)를 통해 일부 옵션 부분을 수정 및 보완하고 CGI 프로그램을 통해서 웹 인터페이스와 연동되게 구현하였다. 본 시스템 구축에 사용된 검색 프로그램들은 HMMPfam, HMMTigr, FingerPRINTScan, HMMSmart, BlastProDom 및 PrositeScan 등이다. 이러한 검색 프로그램들은 개별적으로 해당 목적에 맞게 개발되어 각기 다른 사이트에서 해당 데이터베이스를 대상으로 서비스를 제공하고 있는 것들이다. 이들 프로그램들을 단일 인터페이스 하에서 통합적인 서비스를 제공해 줄 수 있도록 시스템을 구성함으로써 단백질 기능분석의 효과를 극대화 할 수 있도록 하였다. 특히 이러한 검색 프로그램들의 소스를 철저히 파악하여 사용자에게 최적화된 분석옵션 및 분석결과를 제공할 수 있도록 일부 수정 및 보완하였으며, 검색결과는 로컬 사이트에 구축된 데이터베이스와 연동되게 하였다(Fig. 3). 이로 말미암아 사용자는 웹 사이트들을 서핑(Surfing)하는 번거로움 없이 원스톱(One-stop) 서비스로 원하는 모든 분석결과 및 관련 데이터들을 얻을 수 있도록 본 통합 시스템을 구현하였다. 각 CGI 프로그램, 데이터 양식 변환용 프로그램 및 각종 유틸리티들은 C언어, Perl 및 셸 스크립트(Shell Script) 등으로 구현하였다.

InterPro 데이터베이스 및 InterProScan 프로그램

통합 시스템의 데이터베이스 부분 중에서 가장 핵심적인 데이터베이스는 InterPro이다. 이러한 측면에서 비록 상당히 복잡하게 구성되어 있지만 InterPro 데이터베이스의 전체 버전을 다운로드 받아서 로컬 데이터베이스화하였다. InterPro 데이터베이스는 EBI (European Bioinformatics Institute)의 InterPro 2001 공동 프로젝트에 의해 만들어진 것으로, 유일하고 중복성이 없는 단백질 패밀리, 단백질 도메인 및 기능부위 등을 생성함으로써 가장 공통적으로 사용되는 단백질 Signature 데이터베이스들에 대한 최상위 통합계층을 제공한다. 즉, InterPro 데이터베이스는 PROSITE, PRINT, ProDom, SMART, PFam, TIGRFAMs 등의 기존 데이터베이스들을 통합화한 것이다. 이 통합 시스템에서도 InterPro 데이터베이스를 통해 다른 데이터베이스들 사이에 상호참조되게 하였다. 이 연구에서 InterPro 데이터베이스를 로컬 사이트에 구축한 절차를 간략히 살펴보면, 배포되는 XML (eXtensible Markup Language) 양식의 원시데이터와 그에 따른 DTD (Document Type Definition)를 분석하여 어느 부분을 필드로 할 것인지 그리고 테이블 구성을 어떻게 할 것인지 데이터베이스 스키마를 결정하고(Fig. 4 참조), 변환 및 파싱 작업과 데이터베이스 업로드 작업을 통해 생성하였다. Fig. 5에서 나열된 각 필드 및 인덱스에 대해 검색이 가능하며, 검색결과 양식은 "Accession List", "Entry List", 및

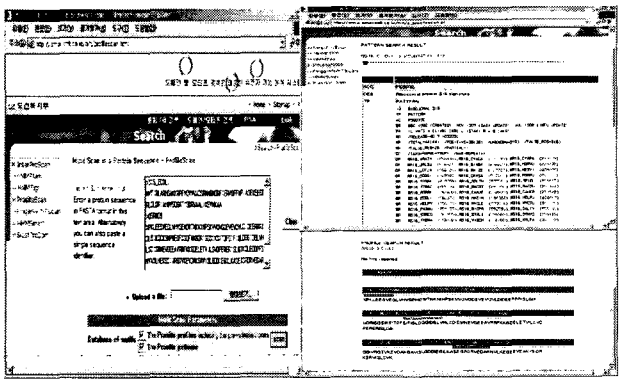

Fig. 3. Web-based user interface of PrositeScan program (left) and its search result (right).

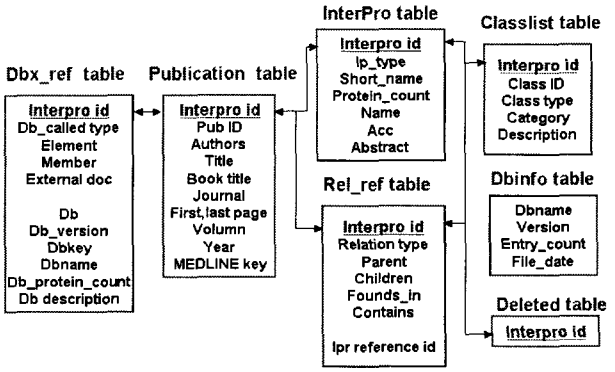

Fig. 4. Schema of InterPro database. It consists of seven tables and the underlined field is primary key, by which corresponding tables are linked.

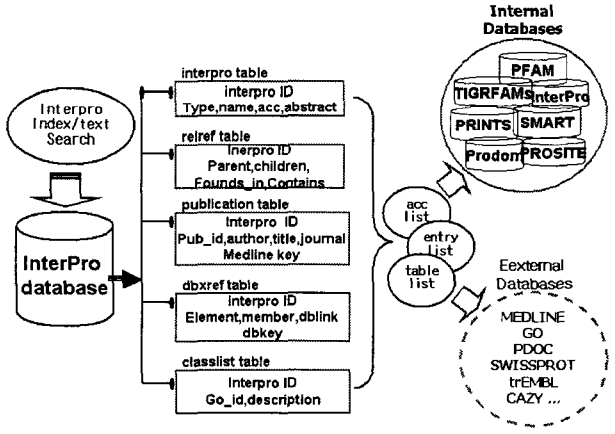

Fig. 5. Index and text search function of InterPro Database. Search result can be represented in three types - accession list, entry list, and table list. InterPro is linked with internal and external databases.

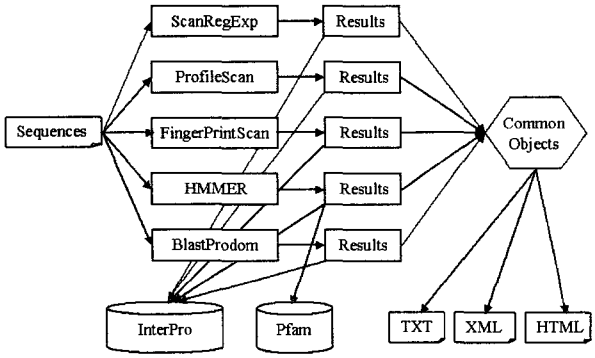

Fig. 6. Overall organization and function of InterProScan. The program internally calls motif search programs (ScanRegExp, ProfileScan, FingerPrintScan, HMMER, BlastProdom) by wrapper function and reports summarized result by means of parsing and reformatting the individual results of motif search programs.

"Table List" 형태로 보여준다. 또한 20여 개 가량의 공개용 외부 데이터베이스들과 하이퍼링크 되게 구현하였다(Fig. 5).

InterPro와 더불어 본 연구개발 시스템의 핵심 요소 중의 하나가 InterProScan[5]이다. 이것은 PERL 기반의 패키지로 InterPro 구성 데이터베이스들에 대한 모티프 및 도메인 통합 검색을 가능하게 해 주는 하나의 도구이다. 이것을 활용하기 위해 일단 PERL 소스 코드 및 내부 구성을 철저히 파악하였다. 분석한 바에 의하면 InterProScan은 하나의 독립적인 검색 프로그램이 아니라 실제 InterPro 구성 데이터베이스들의 개별 검색 프로그램을 래퍼함수(Wrapper function)에 의해서 호출하고 각각의 검색결과를 파싱하고 재포맷하여 일목요연하게 가공하여 사용자에게 제공해 주는 것이다(Fig. 6). 마치 메타 검색 엔진과 유사하다고 생각하면 될 것이다. 그래서 이것들을 세분하고 재구성하여 일반 사용자들에게 최적의 분석 서비스를 제공할 수 있도록 본 시스템 내에 포함시켰다.

결론 및 고찰

세계적 연구의 흐름에 발맞추어 국내에서도 그 어느 때보다도 유전체 연구가 활발히 이루어지고 있다. 정부차원에서도 적극적인 투자와 장려를 아끼지 않고 있을 뿐만 아니라 학계는 물론 산업계에서도 고부가가치 산업으로서의 인식을 토대로 본격적인 유전체 연구가 이루어지고 있다. 이러한 상황에서 당면한 가장 시급한 문제는 대규모 서열결정으로부터 쏟아져 나오는 원시 서열데이터로부터 유용한 유전 정보를 규명하고 발굴하여, 그 결과를 의료, 환경 및 산업공정 분야에 활용함으로써 인류복지에 기여하고 동시에 고부가가치를 추구해야 하는 것이 선결 과제이다. 고전적 실험기법이 아니라 대용량처리 기법으로 신규 유전자 발굴 및 단백질 기능분석을 한다는 것은 궁극적으로 시간과 비용 싸움이라고 할 것이다. 이러한 측면에서 생물정보학은 절대적인 역할을 하고 있고, 향후에는 그 기여도가 더욱 더 커질 것이다. 전체 유전체뿐만 아니라 발현유전정보 데이터에 대한 유용한 유전 정보의 규명에는 기존의 유전 정보를 기반으로 하기 때문에, 유전체 염기서열 및 단백질 수준에서 전체적인 상동성뿐만 아니라 진화론적으로 잘 보존된 특정 영역 간의 상동성에 의존할 수밖에 없는 상황이다. 이러한 측면에서 모티프, 도메인 및 단백질 패밀리 관련 데이터베이스 검색과 분석을 통한 유전자 기능규명 및 분류 시스템은 유전체 연구자들에게 매우 중요한 기반 시스템으로서 역할을 할 것이다.

모티프 및 도메인 검색 대상 데이터베이스들은 내부구조가 상당히 복잡하여 일반 유전체 연구자들이 사용하기에는 많은 어려움이 있다. 또한 내부 생성 알고리즘들이 일반 연구자들이 직관적으로 이해하기에 많은 어려움이 있어 정확하게 이해하고 사용하는 것이 상당히 어려운 문제이다. 현재 개별적으로 검색 서비스를 제공하는 사이트들은 소규모의 단일 분석에만 국한되어 있는 상황으로 실제 연구자들이 추구해야 할 유기적이고 체계적인 분석을 행할 수 없는 상황이라 할 수 있다. 이러한 단일분석을 여러 사이트에서 개별적으로 수행했을 때 그 결과들을 상호연관해서 분석해야 분석결과의 민감도(Sensitivity)와 선택도(selectivity) 등을 높일 수 있으나 현실적으로 연구자들이 그러한 유기적인 검색 및 분석을 할 수 있는 환경이 아니다. 이 연구를 통해 이러한 문제점을 극복하였으며, 웹 인터페이스를 기반으로 하여 기존의 산재해 있는 데이터베이스 정보들을 체계적으로 검색할 수 있는 시스템을 구축하였기에 일반 연구자들의 연구 성과를 극대화 할 수 있을 것이다. 검색 결과들을 파싱 및 재포맷을 통해 종합적이고 체계적이며 유기적인 분석결과를 제공함으로써 새로운 기능을 갖는 유전자 발굴 및 단백질 패밀리별 분류를 통해 유용한 유전정보를 확보하여 국가경쟁력을 높일 수 있을 것으로 여겨진다.

요 약

본 논문은 인터넷 상에 산재해 있는 활용 가능한 다양한 모티프 및 도메인 관련 리소스들을 통합화 함으로써 체계적이고 효율적인 모티프 및 도메인 검색을 할 수 있는 통합 시스템 개발에 대해 다루고 있다. 이러한 시스템을 개발하기 위해 산재해 있는 모티프 및 도메인 관련 개별 데이터베이스의 핵심 부분만을 취합하여 로컬 데이터베이스화하고, 해당 데이터베이스에서 사용되는 개별 분석 도구들을 로컬 서버에 설치하였다. 그리고 다양한 유틸리티 및 CGI 프로그램 등을 통해서 이들 데이터베이스와 분석 도구들을 상호연동시켰고, 단일 웹 인터페이스를 통해 전체적으로 통합하였다. 본 연구에서 개발한 모티프 및 도메인 통합 검색 시스템을 활용한다면, 최종 사용자들은 인터넷 상을 서핑하는데 많은 시간을 낭비하지 않고, 원스톱(one-stop) 서비스를 통해 본인이

하고자 하는 정확한 모티프 및 도메인 검색을 할 수 있어 보다 효율적이고 정확한 단백질 기능분석을 행 할 수 있을 것이다.

감사의 글

본 논문은 2004년도 상명대학교 교내연구비의 지원에 의해서 연구되었음.

참 고 문 헌

1. Attwood, T. K., P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin and C. Zygouri. 2003. PRINTS and its automatic supplement, preprints. *Nucleic Acids Res.* **31**, 400-402.
2. Bailey, T. and C. Elkan. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning Journal* **21**, 51-83.
3. Bateman , A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall and E. L. L. Sonnhammer. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280.
4. Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78-94.
5. Corpet, F., F. Servant, J. Gouzy and D. Kahn. 2000. ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **28**, 267-269.
6. Falquet, L., M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann and A. Bairoch. 2002. The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**, 235-238.
7. Fujibuchi, W. and M. Kanehisa. 1997. Prediction of gene expression specificity by promoter sequence patterns. *DNA Research* **4**, 81-90.
8. Gough, J., K. Karplus, R. Hughey and C. Chothia. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903-919.
9. Haft, D. H., J. D. Selengut and O. White. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371-373.
10. Letunic, I., L. Goodstadt, N. J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R. R. Copley, C. P. Ponting and P. Bork. 2002. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* **30**, 242-244.
11. 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.
12. Schneider, T., G. Stormo and L. Gold. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415-431.
13. Wu, C. H., H. Huang, L. Yeh and W. C. Barker. 2003. Protein family classification and functional annotation. *Comput. Biol. Chem.* **27**, 37-47.
14. Zdobnov, E. M. and R. Apweiler. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848.
15. Zhang, C. and A. K. Wong. 1997. A genetic algorithm for multiple molecular sequence alignment. *Computer Application for Bioscience* **13**, 565-581.