다양한 웹 데이터를 이용한 특정 유기체의 단백질 상호작용 데이터베이스 개발

황 두 성†

요 약

이 논문은 단백질 상호작용 데이터베이스 개발에 관해 기술한다. 개발된 시스템의 특징으로서는 첫째, 생물학자들의 직접적인 실험을 통해 얻어진 단백질 상호작용 및 유전인자 데이터를 제공한다. 둘째, 생물학적으로 관련 있는 다양한 형식의 데이터를 wrapper를 통해 광범위하게 분포된 웹사이트들로부터 추출한다. 셋째, 다양한 웹 데이터들 간의 어휘적, 의미적 이질성을 완화하기 위해 wrapper-mediator에 의한 계층적 모듈 구조를 이용하여 추출된 데이터는 통합 과정을 거친 후, 데이터베이스 저장 및 검색을 가능하게 하였다. 현재까지, 주어진 약 11,500 단백질들에 대해, 생물적으로 의미 있는 데이터를 약 40% 정도 데이터베이스 화 했다. 본 개발된 시스템은 프로티오믹스 연구에서 데이터 분석에 유용할 것으로 기대된다.

Development of an Organism-specific Protein Interaction Database with Supplementary Data from the Web Sources

Doosung Hwang†

ABSTRACT

This paper presents the development of a protein interaction database. The developed system is characterized as follows. First, the proposed system not only maintains interaction data collected by an experiment, but also the genomic information of the protein data. Secondly, the system can extract details on interacting proteins through the developed wrappers. Thirdly, the system is based on wrapper-based system in order to extract the biologically meaningful data from various web sources and integrate them into a relational database. The system inherits a layered-modular architecture by introducing a wrapper-mediator approach in order to solve the syntactic and semantic heterogeneity among multiple data sources. Currently the system has wrapped the relevant data for about 40% of about 11,500 proteins on average from various accessible sources. A wrapper-mediator approach makes a protein interaction data comprehensive and useful with support of data interoperability and integration. The developing database will be useful for mining further knowledge and analysis of human life in proteomics studies.

키워드 : 바이오인포믹스(bioinformics), 프로티오믹스(proteomics), wrapper, mediator, heterogeneous data, 관계 데이터베이스(relational database)

1. 서 론

바이오인포믹스(bioinformics) 분야에서 데이터베이스에 관한 연구 및 서비스가 활발하게 진행되고 있다. 현재, 인터넷 상에서 서비스되는 데이터베이스들에는 비록 상호 참조(cross-reference)가 가능하지만 일관성 없는 데이터 형태(data format) 때문에 상호간의 연관성(association) 처리가 어렵다. 본 시스템은 이러한 어려움을 고려하여 설계되었고 단백질들 간의 상호작용(interaction)과 실제 연구자들의 실험을 통해 밝혀진 단백질 상호작용의 세부사항 및 관련 데이터들을 제공한다. 단백질 상호작용 데이터로서는 단백질, 단백질-단백질 상호작용 그리고 상호작용을 밝히는 세부사항이 될 수 있다. 이러한 데이터를 근거로 단백질 데이터베이스는 생물학적 지식, 그리고 공개된 보충 데이터를 연관시키면 한층 생물학적 의미를 높일 수 있다. 상호작용 데이터 베이스는 단백질의 기능 범주(functional categories), 유전인자(gene), RNA, 유기체(organism) 정보 그리고 참고문헌을 통해 지식 확장이 가능하고 데이터 비교를 통한 검증이 필수적이다. 이러한 보충 데이터는 미래의 프로티오믹스(proteomics) 관련 연구를 위해 가능한 정확성 및 간결성을 유지시키는 것이 필요하다.

바이오인포믹스 분야에서 데이터 실험과 분석은 신생 데이터 발생 및 밝혀진 데이터들 간의 알려지지 않은 사실들을 초래할 수 있다. 이는 기존 데이터에 관한 의미 변형 뿐

† 준 회 원 : Wayne State University, Computer Science Dept.
논문접수 : 2002년 9월 23일, 심사완료 : 2002년 11월 27일

만 아니라 기존에는 없었던 의미가 부여 될 수 있음을 시사한다. 따라서 이들 새로운 데이터 및 그들 관계에 관한 모델화 작업이 수반된다. 거대한 규모의 바이오인포믹스 데이터는 많은 객체들로 구성되어있다. 이들 데이터는 자주 접근되고 갱신되어지며 다양한 부류의 사용자들에 의해 복잡한 질의가 주어질 수 있다 따라서 질의 처리 메커니즘(query processing mechanism)은 간결성을 유지함과 더불어 사용자들에게는 적절한 응답을 제공하여야 한다. 예를 들면 질의 응답이 단백질-단백질 상호작용 혹은 이차원 단백질 순서(sequence)일 경우, 텍스트 형태에 비해 2차원 그래프질의 응답이 보다 유용할 수 있다.

이 논문은 실험을 통해 밝혀진 유기체 초파리(fruit-fly) 단백질 상호작용에 관계된 데이터를 공개된 데이터베이스로부터 추출-연관시키는 방법을 제안한다. 제안된 방법은 초파리의 단백질 상세 정보, 단백질 기능 및 기타 참조사항에 관한 정보를 공개된 데이터베이스로부터 추출한 것과 실제 실험을 통해 얻어진 데이터를 연관시킨다. 추출된 데이터는 생물학적 의미를 유지하면서 단백질 상호작용을 분석하는데 이용된다. 개발된 시스템은 이질적인 데이터 소스(heterogeneous data source) 처리를 위해 wrapper-mediator 구조[10, 11] 기반에서 설계되었다. 또한 추출된 데이터의 일관성 유지를 위해 전역적 스키마(global schema)를 사용한다. 인터넷상에서 서비스되고 있는 데이터베이스로부터 정보를 추출하는 XML, HTML 및 HTML-XML wrapper가 개발되었으며, 사용자 질의는 웹-기반(web-based)에서 수행된다.

이 논문의 구성은 다음과 같다. 2절에서 서비스중인 단백질 상호작용 데이터베이스 및 이질적인 데이터 처리를 위한 관련 연구를 살펴보고, 3절에서 초파리 단백질 상호작용 데이터베이스 시스템 구조 및 설계를 기술한다. Wrapper를 이용한 관련 데이터 추출 기법, 데이터 연관성(association) 및 상호운영(interoperability)을 4절에서 논한다. 5절에서 향후 계획 및 전망에 관해 기술한다.

2. 관련 연구

대부분의 공개된 단백질 상호작용 데이터베이스들은 실험 또는 논문을 통해 알려진 단백질 상호작용 관계 데이터를 보유하고 있다. 이들 데이터베이스들은 데이터 형태, 질의 처리 메커니즘 그리고 시각적 질의 응답 여부에 따라 특징이 정해진다. 데이터 소스만을 고려할 경우, 데이터베이스는 특정 유기체(organism-specific) 데이터베이스(CYPG[3], YPD[6]) 또는 일반적 데이터베이스(general database, BIND[19], DIP[20])로 구분되어진다. 특정 유기체 데이터베이스는 특정 유기체 단백질들간의 상호작용을 다룬다. 질의 처리경우, 시스템은 웹-기반 사용자 인터페이스, 탐색어를 이용한 웹-기반 탐색 메커니즘과 그래픽을 이용한 질의 결과를 제공한다.

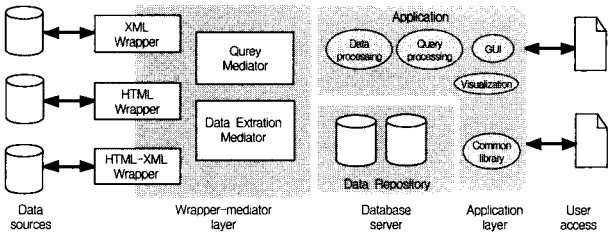
바이오인포믹스 응용분야에서 이 기종 데이터 서비스에 대한 관련 연구로서는 DataFoundary[16], TAMBIS[17], IBM DiscoveryLink[18] 등이 있다. 이 시스템들은 데이터 추출, 변형 및 통합에 있어 wrapper-mediator[10]를 이용하며, 사용자에게 여러 데이터 소스들에 대한 일관성(consistency)을 유지하도록 개발되었다. 메타 데이터 기반(meta-data based) DataFoundary는 추출된 데이터를 mediator을 통해 미들웨어(middleware) 데이터베이스에 저장하여 사용자 질의를 처리한다. TAMBIS는 논리언어 Grail를 이용한 바이오인포믹스 데이터베이스에 대한 지식베이스(knowledge base)를 구성하고 그래픽 사용자 인터페이스를 이용한 데이터 검색 방법을 제공한다. 구축된 지식베이스는 각 데이터 소스의 데이터 모델 표현 및 그들의 데이터 연관성 및 상호운영을 정의한다. 상용화된 DiscoveryLink의 질의 처리기는 사용자 질의를 분해, 해당 웹 데이터에 적합한 질의로 변형 한 후 wrapper를 통해 데이터를 검색한다.

언급된 시스템들과 비교해 볼 때 개발된 데이터베이스는 wrapper-mediator 구조를 기반으로 특정 유기체를 위한 응용 중심(application oriented)에서 설계-개발되었다. Wrapper와 mediator 행위들은 XML 데이터를 기반으로 지시되고 있으며 데이터 통합 및 일관성은 mediator가 다른 웹 소스간의 연관성을 고려하여 wrapper를 작동시킴으로써 이루어진다. 데이터 추출 및 통합이 mediator 중심에서 설계되었고, 초파리 단백질 관련 데이터 분석에 필요한 데이터만을 미들웨어 데이터베이스에 저장하도록 설계되었다.

3. 시스템 구조

개발된 단백질 상호작용 데이터베이스 시스템은 서로 상이한 데이터 형태의 초파리 단백질 관련 데이터 추출 및 통합을 위하여 wrapper-mediator 기법을 기초로 한다[10]. (그림 1)은 개발된 단백질 상호작용 데이터베이스 구조를 보여준다. 이 시스템은 SUN workstation의 관계형 데이터 베이스 Oracle 8i를 사용하고 있으며 사용자 인터페이스, 응용 모듈, mediator 그리고 wrapper 모듈의 계층화 구조(layered structure)로 구성되었다. 여러 관련 소스들로부터의 데이터 일관성을 제공하기 위해 전역적 스키마(global schema)를 사용하였으며, 스키마 설계는 서로 다른 형태의 소스에 존재하는 데이터 연관성(data association)을 이용하여 추출된 데이터를 저장시키도록 하였고 또한 메타 데이터(meta data)를 이용한 추출된 데이터와 관계형 테이블의 세부 데이터에 대한 매핑 기능을 내포하고 있다. Wrapper는 데이터 소스와 HTML 혹은 XML 웹 문서로부터 필요한 데이터를 추출하여 mediator에 제공한다. Mediator는 데이터 매핑, 통합 및 저장에 필요한 일들을 수행하여 일관성 있는 데이터 개념을 유지시킨다. 현재 5 종류의 wrapper 프로그램(GenBank[7], GadFly[2], GeneOntology[5], FlyBase[1], SWISS-PROT[4]

그리고 GeneOntology[5])이 운용되고 있다.



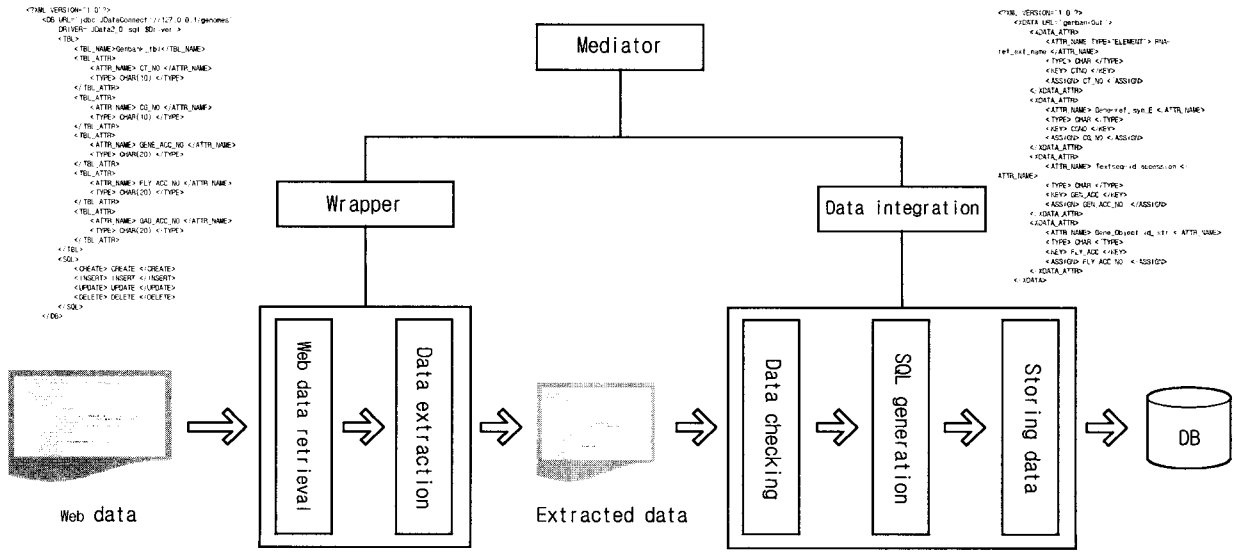
(그림 1) 단백질 상호작용 데이터베이스 시스템 구조

Wrapper와 응용 모듈간의 연결과 데이터베이스 스키마에 대한 데이터 변환(transformation) 및 다중 소스들간의 상호 이용(interoperability)을 제공하는 것이 mediator의 역할이며 이를 위해 고려해야 할 사항들은 다음과 같다 첫째, 인터넷에서 제공되는 이질적 데이터의 검색 및 추출, 둘째, 추출된 데이터의 적절한 변환, 셋째, 전역적 스키마에 따라 단일하게 표현된 데이터들을 통합하는 것이다. 질의 mediator는 전역적 스키마와 사용자 질의간에 서비스를 제공하도록 설계되었다. 또한 데이터 추출 mediator는 전역적 스키마 통합을 위해 설계되어졌으며, 또한 스키마 엔티티(entity), 데이터 관계성(relationship)과 데이터 무결성을 제공한다. 질의 mediator는 데이터 원격접근을 통해 사용자 질의에 응답한다. Wrapper와 mediator 행위를 정의하는 XML 맵 파일(XML map file)은 데이터 추출, 변형 및 저장을 위한 명세(specification)를 정의 한다. 이 파일로부터 mediator는 새로운 wrapper 발생, 데이터 매핑 및 연관 그리고 데이터 저장에 대한 행위를 수행하도록 설계되었다. (그림 2)는 wrapper와 데이터 추출 mediator의 단계별 행위 구조를 XML 맵 파일을 통해 보여준다. Wrapper는 mediator로부터 제공된 XML 환경파일로부터 검색어를 가지고 웹 소스의 HTML 혹은 XML 형태 웹 문서를 검색하고, 필요한 필드를 추출하여 XML 데이터 출력 파일을 생성, mediator에 제공한다. Wrapper의 맵 파일은 검색 경로, 검색어 그리고 추출할 데이터 속성을 정의하게 된다. XML 맵 파일에 의한 wrapper 행위 표현은 XML wrapper 개발을 일반화 할 수 있었다. 현재, 개발된 XML wrapper는 20개 오브젝트들까지 지원하고 있다. 그러나 HTML wrapper는 해당 웹 데이터에 의존하고 있어 일반화가 어렵다. 데이터 추출 mediator는 wrapper로부터 출력된 XML 데이터를 데이터 검사, SQL 생성 및 수행을 한다.

4. 데이터 통합

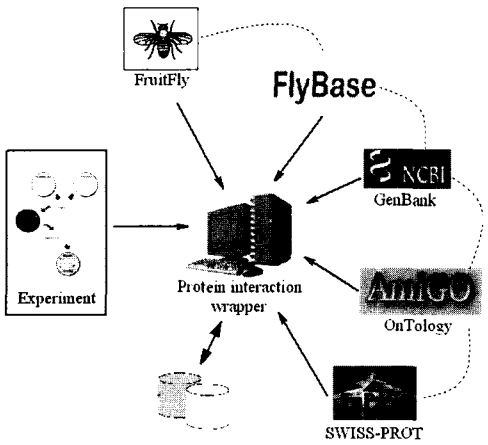
개발한 데이터베이스의 단백질 상호작용 데이터는 초파리 유전자로부터 인코딩된 단백질들간의 모든 가능한 상호반응 실험을 통해 얻어진다. 수집된 데이터들은 단백질 정보의 지식 확장을 위해 각 단백질에 관한 세부사항 들이 웹 또는 텍스트 파일을 통해 데이터베이스에 저장되어진다. 각 데이터는 추출된 소스에 대한 링크정보와 세부 실험데이터를 포함하며, 이들 데이터들은 유전인자(gene)와 단백질에 관한 상세정보를 가지도록 설계되었다.

(그림 3)은 생물학적으로 분산된 웹 데이터베이스간의 초파리 단백질 데이터의 연관성을 보여주고 있다. 여러 소스들로부터 필요한 데이터를 추출하는 wrapper의 각 단계는 다음과 같다. CT(Celera Transcript identification number), CG(Curated Gene identification number) 접근번호(accession number)가 여러 소스들간의 상호작용을 위해 사용된다. CT 번호가 주어지면 GenBank wrapper는 유기체 이름과 FlyBase와 GadFly 접근번호를 제공한다. 다음으로 FlyBase 데이터베이스는 full gene name, synonyms, pheno-



(그림 2) Wrapper-mediator의 데이터 접근-추출-통합 단계

type 정보 그리고 GeneOntology 접근번호를 제공한다 이 GeneOntology 접근번호는 molecular function, cellular component 그리고 biological process의 접근번호를 제공한다. 그리고 FlyBase로 부터 SWISS-PROT 접근번호도 함께 제공된다. GeneOntology로부터 단백질 기능에 관한 데이터를 추출한다. GadFly로부터 mRNA 길이(length)가 얻어지며 SWISS-PROT은 단백질 이름과 synonym들을 얻는다. 각 유전인자와 단백질은 여러 개의 이름과 synonym를 가질 수 있다. 만일 한 단백질이 여러 생물적 프로세스(biological process)에 관련된다면 그 단백질은 여러 생물적 역할을 가지게 된다. XML wrapper가 GenBank, FlyBase 그리고 GeneOntology에서 실행되며 HTML wrapper는 GadFly, SWISS-PROT에서 실행된다. 데이터 검색에 필요한 데이터 링크정보도 함께 데이터베이스에 저장되어, 필요시 상세한 정보까지 제공된다. 접근 가능한 대부분 바이오인포믹스 관련 웹 데이터의 표현은 HTML 기반에서 제공하고 있다. 검색엔진은 파일 시스템이나 데이터베이스에 저장된 적절하게 색인된 데이터를 찾기 위해 정보검색기법을 사용한다. 적합한 검색 결과가 어떻게 나오는지에 상관없이 출력의 형태는 HTML browser를 통해 나타난다. 다음은 wrapper의 개발에 대해서 기술한다.



(그림 3) 단백질 데이터베이스 개발을 위한 분산 데이터베이스와 데이터 연관

- **HTML wrapper**

단백질 상호작용 데이터베이스를 위한 HTML wrapper는 순차적인 텍스트 매칭을 통한 hard-coded 된 형태로 개발되었다. HTML 문서는 순차적으로 검색되어 정보를 추출한다. 따라서 HTML wrapper는 웹 데이터 변화에 상당히 민감하다. 추출된 데이터는 설계된 데이터베이스의 전역적 뷰와 도메인 지식에 의존한다. 전역적 스키마에 변화가 있을 경우 wrapper-mediator의 맴핑 메커니즘을 통해 데이터 재결합이 필요하다. Semi-structured HTML 문서는 적합하지 않은 많은 정보를 수용하고있기 때문에 구문 혹은 의미가 같은 도메인 지식정보가 HTML wrapper에 포함 되어져야 한다. 이점이 HTML wrapper 개발의 일반화를 어렵게 한다. 추출을 원하는 데이터 소스 URL 및 검색어가 주어지면 wrapper는 해당 HTML 문서를 원격 검색, 스트링 형태로 저장 후 추출할 속성을 순차 검색하여 원하는 데이터를 찾는다. 추출 단계는 HTML 문서를 분석하고 wrapper는 XML 형태로 데이터를 출력하도록 개발하였다. HTML wrapper는 GadFly, SWISS-PROT 데이터 추출에 이용되고 있다.

- **XML wrapper**

XML 기반 데이터 포맷과 개념은 인터넷상에 넓게 분포된 이질적인 데이터들의 표준화된 교환을 위해 활발하게 개발되어지고 있다. XML 관련 기법은 일관성 있고 자동적으로 파싱 가능한 포맷으로 데이터를 표현하는 방법들을 제시한다. 바이오인포믹스 데이터 처리에서 XML은 응용 프로그래밍 언어들간의 데이터 이전(migration)에 관한 솔루션을 제공한다. XML를 사용함으로써, 인코딩 혹은 디코딩에 관한 부담없이 바이오인포믹스 응용분야 간의 구조화된 데이터 오브젝트 이동이 가능하다[8]. 따라서 XML은 검색을 훨씬 용이하게 하는 인터넷의 새로운 방법을 제공한다. XML은 여러 태그정보 들이 하나이상의 문서에 적용될 수 있는 기회를 제공한다. 이들 태그는 여러 데이터 요소(element)들을 식별할 수 있게 한다. XML wrapper는 현재 GenBank, FlyBase 그리고 GeneOntology에서 JAVA API DOM을 이용하여 구현되었으며 이들 자바 클래스들은 검색된 XML 문서의 이해를 도울 뿐만 아니라 맵 파일을 통해 데이터 추출을 가능하게 한다. 맵 파일은 모든 데이터 요소들로부터 전부가 아닌 특정부분의 요소들 만을 취했다. 이 맵 파일은 관계 데이터베이스에 매핑 되어지며 XML wrapper는 데이터소스의 DTD가 변할 경우 쉽게 맵 파일의 갱신을 할 수 있다.

- **HTML-XML wrapper**

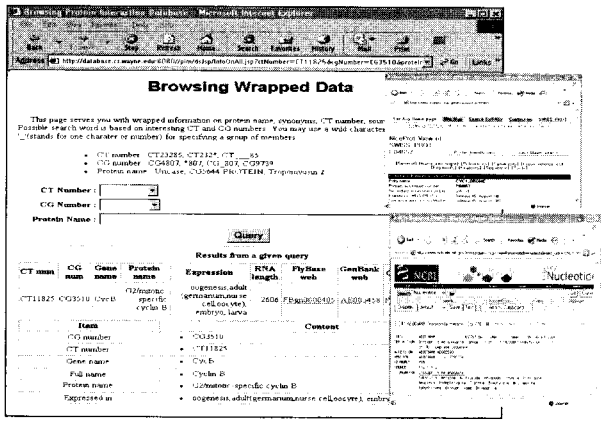
XWRAP Elite는 자동적으로 HTML 문서를 XML 문서로 변화 후 XML 문서로부터 데이터를 추출하는 wrapper 프로그램을 만들어 주는 툴킷이다[15]. 이 툴킷을 이용한 HTML과 XML변환 후 필요한 데이터를 추출하는 wrapper가 테스트되었다. 이 방법은 HTML 문서를 XML 타입으로 변환해주며 간단한 문서를 입력으로 받아 자바로 된 wrapper프로그램을 생성하여 HTML 문서를 구조화된 XML문서로 변환시킨다. 다음은 XWRAP Elite의 시스템 흐름에 관해 기술하고 있다. 첫 번째로 XWRAP Elite는 주어진 샘플을 HTML tree 형태로 파싱한 다음 불필요한 사항들을 걸려낸 후, 의미 있는 정보만을 추출한다. 둘째로 XWRAP Elite는 오브젝트들을 element 단위로 분할한 후, 세번째 단계에서는 분할된 element들을 논리적인 그룹으로 결합한다. 최종적으로 모든 것들이 well-formed된 문서로 패키지화 된다.

XWRAP Elite의 단점은 오브젝트의 수가 최소 5에서 최대 50까지로 제한 된다는 것이다. GenBank에서의 오브젝트수는 50이상 이기때문에 이 툴은 적절하게 이용될 수 없다.

<표 1>은 추출된 데이터의 비율을 보여주고 있다. 현재 데이터베이스는 주어진 11,595 단백질 중 40%에 해당하는 단백질, 약 5000정도의 단백질에 대한 기능 정보, 그리고 약 11,000 단백질에 관한 유전인자 정보를 보유하게 되었다. 이런 사실은 유전인자 정보가 단백질 정보 보다 훨씬 풍부함을 시사한다. 공식적 데이터베이스가 증가와 더불어 wrapper는 더욱 많은 정보를 추출하게 될 것으로 기대된다.

〈표 1〉 공개된 데이터베이스로부터 추출된 데이터 비율

내 용	데이터 수	비율	데이터 소스
Gene information	11,595	100.0%	GenBank, FlyBase
Gene synonym	3,955	34.1%	GenBank, FlyBase
Molecular function	4,980	43.0%	GenBank, GeneOntology
Biological process	3,255	28.1%	GenBank, GeneOntology
Cellular component	3,502	30.2%	GenBank, GeneOntology
Protein name	3,235	27.9%	GenBank, Swiss-Prot
Protein synonym	2,253	19.6%	GenBank, Swiss-Prot
mRNA length	11,595	100.0%	GenBank, GadFly



(그림 4) 데이터베이스 질의 처리의 예

개발된 데이터베이스 시스템은 웹-기반 질의 처리를 지원한다. 웹 질의 페이지는 사용자로부터 질의 실행에 필요한 인자를 입력하도록 유도하여 올바른 SQL 문자을 생성 실행 후 검색된 데이터를 사용자에게 제공한다. (그림 4)는 CT 번호 "CT11825" 데이터 검색 결과를 보여 주고 있다. 검색된 데이터가 웹 링크이면 원격 데이터에 대한 웹 페이지를 참조할 수 있다. 그림에서 해당 CT에 대한 데이터베이스 정보와 관련 GenBank, SWISS-PROT 웹 페이지를 같이 검색하고 있다.

5. 결 론

이 논문은 wrapper-mediator 기법을 적용한 단백질 상호작용 데이터베이스 구현에 관하여 논의하였다. 단백질 데이터베이스는 생물학적 지식과 그 지식에 근거한 보충 데이터 간의 연관성이 필요하며, 데이터베이스에 저장된 데이터의 검증이 요구된다. 단백질 상호작용 데이터베이스는 분석을 위해 데이터의 정확성 및 간결성이 요구된다. 공개된 데이터베이스로부터 단백질 관련 데이터를 추출-저장함에 있어서, 데이터 이질성은 데이터 소스의 형태에 따라 설계되어지는 wrapper의 필요성을 유도하였다. 이 개발된 wrapper들은 HTML과 XML 문서 모두를 처리할 수 있도록 독립적으로 설계-개발되었으며 실험실에서 발생한 단백질 상호작용 데이터의 생물학적 의미를 높이는데 이용되었다. 논의 된 응용 프레임은 다른 유기체의 단백질 상호작용 데이터베이스 개발에 유용할 것으로 기대된다. 현재, 개발된 초파리의 단백질 상호작용 데이터베이스는 단백질 기능 분석 및 예측 및 상호작용 경로(interaction pathway) 분석에 이용되고 있다. 개발된 시스템은 곧 인터넷상에서 이용 가능하게 될 것이다.

참 고 문 헌

[1] The FlyBase Consortium, The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Research* **30**, http://flybase.org, pp.106-108, 2002.

[2] GadFly : Genome Annotation Database of Drosophila, http://flybase.bio.indiana.edu/annot/.

[3] Munich information center for protein sequence, http://mips.gsf.de/proj/yeast/.

[4] SWISS-PROT, http://www.expasy.ch/sprot/sprot-top.html.

[5] GENE ONTOLOGY CONSORTIUM, http://www.geneontology.org.

[6] IncyteGenomics For Life, http:// www.incyte.com/proteome/YPD.

[7] GenBank, http://www.ncbi.nlm.nih.gov/GenBank.

[8] W3C World Wide Web, http://www.w3.org.

[9] Frederic Achard, Guy Vaysseix and Emmnauel Barillot, "XML, Bioinformics and data integration," Bioinformics Review, Vol.17, No.2, pp.115-125, 2001.

[10] Gio Wiederhold and Michael Geneserecth, "The Conceptual Basis for Mediation Services," IEEE Expert, Intelligent Systems and their Applications, Vol.12, No.5, 1997.

[11] Hector Garcis-Molina, Yannis Papakonstantinos, Dallan Quass, Anand Rajaraman, Jeffrey Ullman, Jennifer Widom and Vasilis Vassalos, "The TSIMMIS Approach to Mediation : Data Models and Languages," Journal of Intelligent Information Systems, Vol.8, No.2, pp.117-132, 1997.

[12] Mary Tork Roth and Peter Schwarz, "A Wrapper Architecture for Legacy Data Sources," IBM, Technical Report

RJ10077, 1997.

[13] Merja Ek, Heli Hakkarainen, Pekka Kilpelainen, Eila Kuikka and Tommi Penttinen, "Describing XML Wrappers for Information Integration," University of Kuopio, Technical Report A/2001/2, 2001.

[14] Vincent H Guerrini and David Jackson, "Bioinformics and Extended Markup Language(XML)," Online journal of Bioinformics, Vol.1, No.1, 2000.

[15] Wei Han, David Buttler and Calton Pu, "Wrapping Web Data into XML," SIGMOD Record, Vol.30, No.3, pp.33-45, 2001.

[16] T. Critchlow, K. Fidelis, M. Ganesh, R. Musick and T. Slezak, "DataFoundry : Information Management for Scientific Data," Processdings of IEEE Advances in Digital Libraries, 2000.

[17] C. A. Goble R. Stevens, G. Ng, S. Bechhofer, N. W. Paton, P. G. Baker, M. Peim and A. Brass, "Transparent access to multiple Bioinformics information sources," IBM Systems Journal, Vol.40, No.2, pp.532-551, 2001.

[18] IBM Life Science Solution Team, IBM Life Science Solutions : Turing Data into Discovery with DiscoveryLink, Redbooks, 2002.

[19] Gary D. Bader, Ian Donaldson, Cheryl Wolting, B. F. Francis Oullette, Tony Pawson, and Christopher W. V. Hogue, "BIND-The Biomolecular Interaction Network Database," Nucleic Acids Research, Vol.29, No.1, pp.242-245, 2001.

[20] Ioannis Xenarios, Danny W. Rice, Lukasz Salwinski, Marisa K. Baron, Edward M. Marchotte and David Eisenberg, "DIP : the Database of Interacting Proteins," Nucleic Acids Research, Vol.28, No.1, pp.289-291, 2000.

황 두 성

e-mail : dhwang@cs.wayne.edu

1985년 충남대학교 계산통계학과 졸업 (학사)

1990년 충남대학교 대학원 계산통계학과 (석사)

1990년~1991년 국토개발연구원 연구원

1991년~1998년 전자통신연구소 연구원

1998년~현재 Wayne State University, Computer Science Dept.(박사과정)

관심분야 : 데이터 마이닝(data mining), 머신 학습(machine learning), 바이오인포믹스(bioinformics)