당뇨 연구를 위한 웹기반 통합 데이터베이스 시스템 구현

김재희, 류근호
충북대학교 전자계산학과
e-mail : jaehei@knfc.co.kr, khryu@dblab.chungbuk.ac.kr

Web-Based Integrated Database system Implementation for Diabetes Research

Jae-Hee Kim, Keun-Ho Ryu
Database / Bioinformatics Laboratory, Chung-buk University

요 약

오늘날 생물학 데이터베이스 시스템은 끊임없이 증가하고 복잡하게 연결되는 데이터를 처리해야 할 필요성과 데이터양의 증가만큼이나 빨리 성장하는 사용자들의 요구에 부응해야 하는 필요성에 직면해 있다. 이 논문에서는 기존의 생물학 데이터베이스 시스템의 특징을 살펴본 후, 현재 당뇨 관련 데이터베이스가 존재하지 않으므로 당뇨 연구를 위한 포괄적인 정보 제공과 사용의 편의를 제공하기 위하여 생물학 관련 데이터베이스를 교차 참조한 당뇨 연구용 데이터베이스를 설계하였다.

본 논문에서 설계한 데이터베이스는 Genetic Information, Protein Information, In Silico digestion, 그리고 Chemical Information 4개의 메뉴로 구성하였다. Genetic Information과 Protein Information은 Cross-Reference를 통한 관련 데이터베이스와 연결시켰고, Protein Information에서 PDB 코드가 존재할 경우 3차원 분자 구조를 제공한다. 아울러 단백질 동정시에 활용할 수 있는 선택된 효소처리 후의 펩타이드의 이론적 질량값을 계산하도록 구현(In Silico Digestion)하였으며, 당뇨 관련 주요 단백질의 화합물들의 구조를 제공하였다.

1. 서론

당뇨병(Diabetes)은 혈중에 있는 당분이 인슐린 분비의 부족이나 인슐린의 작용 및 기능 부족으로 인해 에너지로 사용되지 못하고 혈액에 남아 있게 되는 질병으로 원인으로는 식생활의 서구화, 비만증, 운동부족, 스트레스 등을 들을 수 있다. 현재 미국, 유럽 및 일본과 같은 선진국에서는 비만 및 당뇨병의 증가로 인한 사회 경제적인 파장이 엄청나며, 한국의 경우도 경제적 발전으로 인해 당뇨병 환자가 급속도로 증가하여 2025년경에는 3억명 이상의 환자가 발생할 것으로 예측된다.[1,2] 결국 당뇨병은 특정 요인에 국한되어 연구하기 보다는 포괄적인 접근이 필요하고, 이전에 진행된 연구들의 풍부한 자료 수집이 필요하다.

본 논문에서는 생물학 관련 데이터베이스를 교차 참조하여 통합된 데이터베이스로부터 당뇨 연구를 위한 Small 데이터베이스를 구축함으로써 효율적으로 당뇨 연구자들이 접근할 수 있고, 자료 검색 시간의 단축으로 보다 빠른 연구 자료를 제공하고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 바이오 관련 연구 및 바이오 정보 통합 기술에 대해 설명한다. 제 3장에서는 당뇨 관련 데이터베이스 설계에 대해 설명한다. 제 4장에서는 본 시스템의 구현에 대해서 설명한다. 제 5장에서는 결론을 기술하도록 한다.

2. 웹 기반 바이오 관련 연구

현재 존재하는 바이오 데이터베이스는 DNA 염기 서열 정보, 아미노산 서열 정보가 주류를 이루고 있다. 일반적인 바이오 데이터베이스에는 GenBank[3], EMBL[4], PRI-International[5],

Swiss-prot[6], Protein Data Bank[7]등이 있다. 여기에서 대표적인 GenBank에 대해 살펴본다.

2.1 GenBank

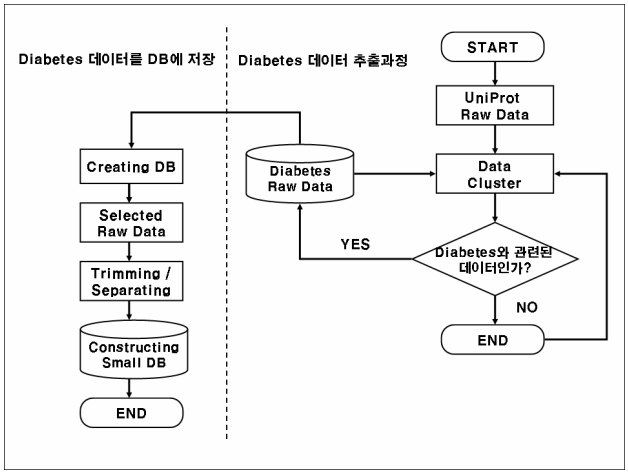
GenBank는 미국 NIH(National Institute of Health)의 후원으로 운영되는 최대의 유전자 데이터베이스이다. 이곳은 NCBI라는 곳에서 운영을 도맡아 하고 있으며 유럽 EMBL과 일본 DDBJ와 자료교환 및 공유를 하고 있다. 또한 GenBank는 최대의 정보를 가지고 있고 다른 유전자 서열과의 상동성 검색이 가능하며 데이터베이스로부터 특정 유전자 서열을 불러 낼 수 있으며 Entrez(Integrated database retrieval system)[8]으로부터 Nucleotide, Protein database를 관련 문헌과 함께 검색할 수 있다. 또한 짧은 염기 서열에서 genomic 서열과 OMIN에서 제공하는 Phenotypic description, MMDB(Molecular Modeling database)를 통한 단백질 구조도 서열 정보와 함께 링크되어 있다.

2.2 바이오 정보 통합 기술

기존에 구축된 데이터베이스 시스템들은 각 기관별로 독자적으로 구축되었기 때문에 사용된 소프트웨어나 저장형태가 다양하고 지리적으로 분산되어 있다. 따라서 다양한 형태로 저장된 정보를 효과적으로 접근 및 관리할 수 있는 바이오 정보 통합 관리 기술 개발이 필요하다. 바이오 정보 통합 기술에는 세가지로 나눌 수 있다. 첫째 응용 프로그램 수준으로 통합하는 기술은 현재 가장 많이 쓰이는 방식이지만, 새로운 소스 추가/변경이 어렵고, 사용할 수 있는 인터페이스도 제한적이다. 둘째 데이터웨어하우스 기술[9]을 이용한 물리적 통합은 성능 면에서 다소 장점이 있지만 데이터를 중복해서 저장함으로써 저장 비용 증가, 관리 비용 증가, 원 정보 저장소의 특수 검색 기능 손실 등의 단점이 있다. 셋째 가상 데이터베이스 기술을 이용한 링크 기반 통합 기술은 비교적 손쉽게 구축할 수 있어서 다른 것에 비해 비교적 많이 활용되는 방법이다. 본 논문에서는 링크 기반 통합 기술에 중심을 두는 웹 기반 통합 검색 시스템을 설계 구현한다.

3. 당뇨병 관련 데이터베이스 설계

현재 2007년 4월을 기준으로 GenBank 데이터베이스에는 당뇨관련 핵산서열이 39,909개, 단백질 서열이 16,491개 등록되어 있으며, UniProt 데이터베이스에는 단백질 서열이 약 600개 정도 등록되어 있다. 이들 모두를 다운로드 받았으며, 서열 데이터의 형태는 텍스트 파일이다. 이 텍스트 파일의 전체적인 크기는 약 100GB에 이르므로 하나 하나 입력하는 것은 불가능하다. 이에 다운받은 염기서열 텍스트 파일은 분석해 파싱한 다음, 그 결과로 나온 데이터를 바탕으로 nrDB(non-redundant)에서 Nucleotide Sequence와 자동 연결, 입력하는 프로그램을 작성하였다. (그림 1)은 본 프로그램의 알고리즘을 도식화하고 있다. 이 프로그램은 크게 두가지 기능을 제안한다. 첫 번째 기능은 GenBank나 UniProt에서 다운로드한 데이터에서 “Diabetes” 키워드가 포함된 데이터만을 선별하여 추출하는 기능이다. 두 번째 기능은 "Diabetes" 데이터 블록으로만 이루어진 데이터베이스를 각 테이블 형식에 맞도록 데이터를 잘라서 구조화된 Small 데이터베이스를 만드는 기능이다. 이 프로그램은 당뇨 관련 정보 뿐 아니라 다른 주제에 대해서도 선택적으로 데이터베이스를 구축할 수가 있으며, 사용자 정의에 의한 2차 Small 데이터베이스이므로 기존의 통합 데이터베이스에 비해 특정 분야의 연구자들에게 빠른 검색 서비스를 제공할 수 있다.



(그림 1) 데이터 추출 알고리즘

4. 구현

본 논문에서 설계한 당뇨 관련 데이터베이스는 Genetic Information, Protein Information, In Silico Digestion, 그리고 Chemical Information으로 구성하였다.

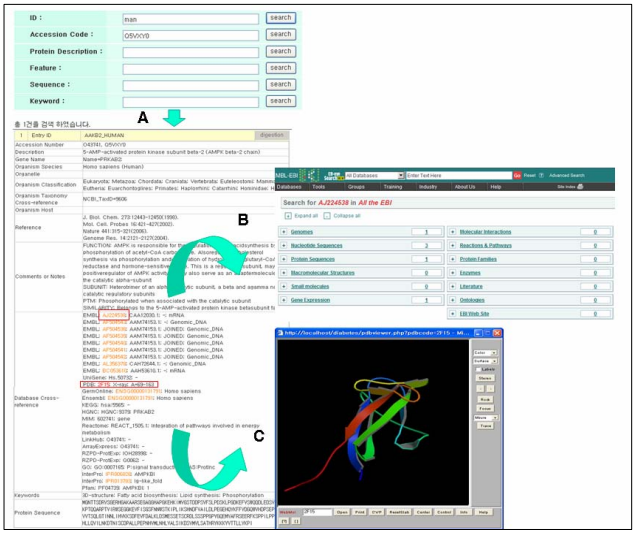
웹 서비스를 검색하는 시스템 환경은 Apache를 이용해 구현하였고, User Interface부분은 PHP를 이용해 구현하였고, Database는 Mysql을 이용하였고, 데이터 파싱은 Perl5.88을 이용하였고, 3D Structure

viewer는 JMOL, WebMOL을 통해 구현 하였다.

4.1. Genetic and Protein information 검색

본 논문에서 지원하는 질의문의 형태는 Pattern Matching과 'AND' Boolean Operator의 결합으로 연결되어 모호한 다수의 질의어에 대한 필터 기능으로 사용 할 수 있도록 설계하였다. 이는 '통합검색'의 개념과 같으며, 데이터베이스 자체가 당뇨라는 주제에 일치하는 정보만을 추출해 구성하였기 때문에 'garbage out' 현상이 매우 적고, 사용이 간편하고, 결과에 대한 정보의 유용성이 매우 높다.

또한 본 논문은 PDB 코드가 존재할 경우 JMOL을 이용한 3차원 분자 구조를 실시간으로 보여주어 연구자들이 결과 정보를 이해하는데 도움을 주고, 3차원 구조를 찾아보는 시간과 노력을 절감하였다. 그리고 검색 결과 관련된 유용한 데이터들은 Cross-Reference-Link를 사용하여 EMBL, GO, Swiss-Prot, PDB, KEGG와 같은 주요 데이터베이스에 연결하였다. (그림 2)의 A는 ID에 "man" 키워드를 입력하고 Accession number에 "5QVXY0" 키워드를 입력하여 두 개의 질의어에 해당하는 정보를 검색한 화면이다. B는 검색 결과 중 유용한 관련 된 데이터들을 Cross-Reference-Link를 사용하여 관련된 데이터베이스에 연결한 화면이다. C는 JMOL을 이용해 3차원 분자 구조를 보여주는 화면이다.



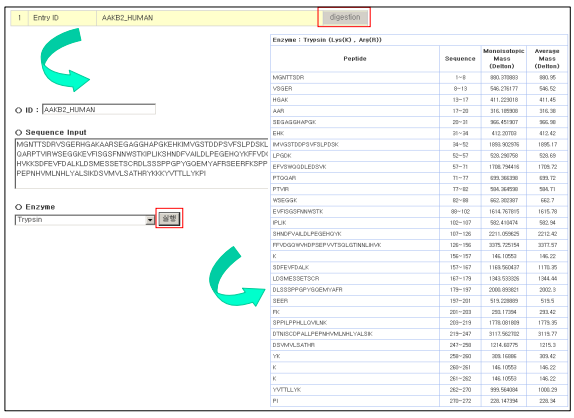
(그림 2) 정보검색 구현

4.2. In Silico Digestion

본 논문에서는 당뇨와 관련한 특정 단백질이 효소에 의해 잘려진 펩타이드 조각들의 질량값을 계산하는 모듈을 제공함으로써 당뇨관련 단백질 동정시에 참고치(이론적 질량값)로 활용하고자 한다. 그리고 이 모듈은 protein information에서 선택한 단백질의 서열이 바로 연동 되도록 구현 하였다.

단백질 동정이란 무엇인가 간략히 살펴보면, 인간의 유전자는 3만개 정도에 불과하나 이들이 만들어 낸 단백질 중에서 인체 내에서 기능을 가지는 단백질은 3만 ~ 10만개 정도 존재하며, 변형된 것까지 합하면 천문학적인 숫자의 단백질이 존재한다. 따라서 유전자의 산물인 단백질을 대상으로 이들을 대량 분석하고, 상호 관계 지도를 작성하며, 구조 분석을 통해 특정 단백질의 기능을 밝히는 것은 매우 중요하다. 이와 같은 학문을 프로테오믹스(단백질체학)이라 하는데 크게 두 가지로 나눌 수 있다.

Top-down 분석방법은 단백질 자체를 질량 분석기에 넣어 에너지를 가하여 조각을 내고 각각의 조각 무게를 측정하여 단백질 서열을 확인하는 동정법이고, Bottom-up 분석방법은 효소를 이용해 단백질의 특정 아미노산을 절단하여 여러 펩타이드로 나눈 후 이 펩타이드들을 질량 분석기로 분석하여 구해진 질량 값들을 가지고 단백질이 무엇인지 알아내는 방법이다. 이 과정에서 얻어진 질량 값들은 단백질 데이터베이스와 비교를 통해 각각의 단백질로부터 이론적으로 계산된 펩타이드의 질량값과 비교를 통해 단백질을 동정하는 방법이다. 본 논문에서는 당뇨와 관련한 특정 단백질의 확인 또는 변형에 대한 정보를 얻기 전 미리 이론적인 계산값을 얻음으로써 참고치로 활용하고자 한다. (그림 3)은 Entity ID가 "AAKB2_HUMAN" 인 단백질 동정 정보를 확인한 화면이다.



(그림 3) 단백질 동정 구현

4.3. 화합물 정보제공

본 논문에서는 당뇨관련 신약 개발에서 활용할 수 있도록 기존에 연구된 주요 단백질의 화합물 구조를

제공하였으며, 신약 개발에서는 타겟 단백질과 결합하는 화합물의 골격구조를 정하는 일은 매우 중요하며, 이들은 기존에 연구 되어진 화합물을 바탕으로 골격을 정하거나, 가상 스크리닝을 통해 선도물질을 정한다. 단백질의 활성부위에 잘 결합할 수 있는 화합물을 찾는 과정은 분자 설계를 통해 이루어지는데 단백질에 화합물을 컴퓨터상으로 가상으로 결합시키는 과정을 "Docking"이라 하며, 이를 통해 결합에너지 계산, 시뮬레이션이 가능하고, 새로운 화합물의 De novo design이 가능하다. 본 논문에서 구축된 화합물 데이터베이스는 기존에 연구되어진 당뇨 관련 주요 화합물 구조에 대한 정보를 제공함으로써 새롭게 연구하거나 관심이 있는 연구자들에게 참고자료로 유용하게 활용될 수 있다.

5. 결론

본 논문에서는 당뇨 관련 연구자들이 당뇨 관련 정보를 검색할 수 있도록 지원하기 위해, 생물학 관련 데이터베이스에서 당뇨 관련 데이터만을 추출해 핵산서열 및 단백질 서열 정보 등을 링크 기반으로 통합하였다.

본 논문에서 개발한 시스템은 통합 메타 정보 관리기, 질의 관리기, 연동기, 결과 관리기 등 4개의 블록으로 구성하였고, 당뇨 관련 연구에 필요한 방대하고 이질적인 데이터들을 한 곳에 모아 통합 검색을 지원함으로써, 연구자들이 원하는 정보를 쉽게 획득하고, 데이터 검색 시간을 단축할 수 있도록 하였다.

본 논문에서 개발된 시스템이 다른 통합 정보 시스템과 차별성을 갖는다면 언제라도 관심 주제만 바꾼다면 사용자 정의 2차 데이터베이스를 구축할 수 있으며, 데이터베이스를 특정 주제로 소형화 시켜 결과 정보의 정확성, 신속성의 장점을 모두 갖춘 데이터베이스라는 점이다. 또한 Cross-Reference-Link를 통해 관련된 데이터베이스와 연결하였고, PDB 코드가 존재할 경우 3차원 분자 구조를 제공하고, 단백질 확인이나 변형에 대한 정보를 얻어 참고치로 활용될 수 있는 단백질 동정 정보와 당뇨 관련 주요 단백질의 화합물들을 제공한다.

본 논문은 당뇨 관련 연구자들에게 필요한 방대하고 이질적인 데이터들은 한곳에 모아 통합 검색을 지원함으로써, 생명공학 관련 기초 연구나 신약물질 개발에 필요한 제반 지식을 빠르게 획득할 수 있게 되어 급속도로 성장하고 있는 생명공학 분야의 연구에 도움을 줄 것으로 기대된다.

참고문헌

[1] Park Y, Yoo K, Lee H, Kim Y, Koh CS, Shin Y,Min H. 1995. Prevalence of diabetes and IGT in Younchon county. Diabetes Care 18: 545-548

[2] Son HS, Song KH, Han JH, Lee JM, Youn KH, Kang MI, Cha BY, Lee KW, Son HY, Kang SG.1994. The prevalence of diabetes mellitus and its relation to body mass index in Koreand subjects. Diabetes Suppl 1:07A0990037

[3] Abola,E.E., Sussman,J.L., Prilusky,J. and Manning,N.O. (1997) Protein Data Bank archives of three-dimensional macromolecular structures. Methods Enzymol., 277, 556-571.

[4] Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor, M. and Sterk,P. (1998) The EMBL nucleotide sequence database. Nucleic Acids Res., 26, 8-15.

[5] Barker,W.C., Pfeiffer,F. and George,D.G. (1996) Superfamily classi-fication in PIR-International protein sequence database. Methods Enzymol., 266, 59-71.

[6] Bairoch,A. and Apweiler,R. (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res., 26, 38-42.

[7] Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) GenBank. Nucleic Acids Res., 26, 1-7.

[8] Entrez online documentation : http://www.ncbi.nlm.nih.gov/Database/index.html

[9] Stein,L. Integrating Biological Databases. Nature Reviews-Genetics Vol4. 337-345. 2003