분산된 생물정보 데이터베이스의 통합 검색 시스템 연구

윤홍원°
신라대학교 컴퓨터정보공학부

Integrated Information Retrieval System from Distributed Biological Database

Hongwon Yun
Dept. of Computer Engineering, Silla Univ.

요 약

분자 생물학의 발전으로 염기 서열, 단백질 서열, 지놈 서열 등의 서열 데이터베이스와 단백질 3차 구조를 제공하는 구조 데이터베이스 등이 구축되어서 웹을 통해 많은 정보를 제공하고 있다. 전세계적으로 분산되어 있는 다양한 생물정보 데이터베이스의 효율적인 검색을 위해서 통합 검색 시스템의 개발이 필요하다. 이 논문에서는 전세계의 생물정보 데이터베이스의 개발 현황을 보이고 분산되어 있는 생물정보 데이터베이스로부터 통합 검색을 위한 생물정보 통합검색시스템 GenPlus를 제안하였다. 제안한 GenPlus에서는 염기 서열, 단백질 서열, 그리고 키워드를 이용한 서열 정보, 구조 정보, 완전한 지놈 정보, 그리고 문헌 정보의 통합 검색을 제공한다.

1. 서론

최근 분자 생물학 기술의 발달과 빠르게 진행되고 있는 지놈 프로젝트를 통해 방대한 양의 유전자 서열 정보와 새로운 형태의 생물학 정보들이 쏟아져 나오고 있다. 염기 서열, 단백질 서열, 단백질 3차 구조, 지놈 서열, 대사 경로 등의 다양한 생물정보 데이터베이스가 구축되어서 웹을 통해서 많은 정보가 제공되고 있다.

생물학 관련 데이터베이스들은 그 초점을 어디에 두는가 따라 여러 가지 방식으로 데이터를 가공하고 표현한다. 염기서열, 단백질 서열 정보를 제공하는 GenBank, 서열의 기능 정보가 정리되어 있는 SWISS-PROT 등이 있으며, 단백질 3차 구조에 관한 자료는 SCOP과 CATH 에 잘 정리되어 있다. 생화학 대사 경로 데이터베이스로 잘 알려진 KEGG와 WIT 등이 있으며 전체 지놈 서열이 밝혀진 서로 다른 종의 상호 비교 결과는 COG, MBGD, KEGG의 데이터베이스에 정리되어 있다. 외형과 분자 생물학적인 증거를 바탕으로 개체를 분류한 정보는 NCBI taxonomy 데이터베이스와 Tree of life 등이 가지고 있다[1,2].

이러한 생물정보 데이터베이스는 전세계에 분산되어 있으므로 효율적인 정보 검색을 위한 방법이 필요하다. 보기를 들어보면 NCBI의 GenBank와 PIR은 워싱턴, HGD는 메릴랜드, SWISS-PROT은 스위스, DDBJ는 일본에 위치하고 있으며 이러한 생물정보

데이터베이스의 수가 500여개에 이른다[3]. 전세계에 분산되어 있는 이질적인 생물정보 데이터베이스의 통합 검색을 위한 연구가 태동 단계에 있다. 지금의 생물학자들이 DNA서열 검색, RNA 서열 검색, 단백질 기능 검색 등을 단계별로 밟아 가는 과정은 불편하고 시간의 낭비를 가져오고 있다. 전세계에 분산되어 있는 생물정보 데이터베이스에서 한번의 질의를 통해서 원하는 정보를 얻을 수 있는 데이터베이스와 툴이 필요하다.

본 연구에서는 생물정보 데이터베이스의 현황을 보이고, 한번의 질의로 전세계에 분산되어 있는 생물정보 데이터베이스에서 원하는 정보를 얻을 수 있는 통합 검색 시스템을 제안한다. 본 연구에서 제안하는 생물정보 통합 검색 시스템에서는 키워드나 접근기호 등을 포함하는 하나의 질의로 원하는 정보를 얻을 수 있으며 서열 데이터베이스, 대사경로 데이터베이스, 지놈 데이터베이스, 문헌 데이터베이스가 연결된다.

2. 관련 연구

NCBI에서 구축한 Entrez는 분자 생물학 검색 시스템은 염기나 단백질 서열 데이터베이스, 분자 모델링 3차원 구조 데이터베이스, 지놈과 맵 데이터베이스와 문헌들을 검색할 수 있다. 사용하기가 쉽지만 제한된 검색 정보를 제공한다. 검색은 하나의 데이터베이스에서 시작하고 질의에 해당하는 레코드들을 결과로 보여준다. SRS는 영국의 EBI에서 개발한 80여개의 생물학 데이터베이스에 대한 단일 인터페이스이다. 서열, 대사 경로, 단백질 구조, 지놈, 맵핑, 돌연변이 등의 여러 데이터베이스가 포함되어 있다. 검색 데이터베이스들이 많이 있지만 인덱싱이 잘 되어 있어서 검색에 걸리는 시간이 짧다. DBGET/LinkDB는 일본 교토 대학교에서 개발하여 GenomeNet을 통해서 이용할 수 있는 통합 데이터베이스 검색 시스템이다. DBGET은 약 20개의 데이터베이스를 한번에 하나씩 검색할 수 있도록 한다. 하나의 데이터베이스에서 질의를 하면 DBGET은 결과물과 함께 관련 정보들의 링크도 제공한다. DBGET은 SRS나 Entrez에 비해서 간단하지만 검색방법이 제한적이다[6].

미국, 유럽, 일본에서 개발한 생물정보들은 서로 공유되고 있으나 Entrez, SRS, 그리고 DBGET은 자체 데이터베이스를 중심으로 검색을 지원하고 있다.

분산되어 있는 데이터베이스에 접근하는 방법 두 가지를 살펴보면 다음과 같다. 원격지에 위치한 다수의 데이터베이스에 질의를 보내서 결과를 받고, 받은 결과를 분석하고 재정돈하여 마지막 결과를 사용자에게 돌려주는 접근 방법을 데이터베이스 퓨전 또는 데이터베이스 머지라고 부른다[4]. 웹 메타-서치 엔진인 프로퓨전과 메타크롤러가 방법을 쓰고 있으나 횡단-참조를 해야 하는 본 시스템에는 적합하지 않다. 다른 접근 방법으로는 모든 질의를 수행하여 모든 정보를 한 곳에 모으는 스파이더나 데이터 파일 교환을 쓸 수 있으나 방대한 양의 생물정보를 전송하여 저장하기에는 적합하지 않다. 또한, 서로 다른 데이터베이스로부터 전역 스키마를 만들기가 어렵다는 문제점을 가지고 있다[5].

본 연구에서는 메타-서치와 데이터를 한 곳에 모으는 방법을 혼합한 하이브리드 접근 방법을 고려하고 있다. 분산되어 있는 각각의 원래 데이터베이스에 접근할 수 있는 매칭 엔트리를 유지하고 또한 URL 포인터를 이용하여 통합 검색을 지원한다. 통합 검색의 영역은 서열 데이터베이스, 대사경로 데이터베이스, 지놈 데이터베이스, 문헌정보 데이터베이스를 포함한다.

3. 생물정보 데이터베이스 분류 및 현황

3.1 인포바이오젠 데이터베이스 목록

Infobiogen에서 유지하고 있는 생물정보 데이터베이스 (http://www.infobiogen.fr/services/dbcat) 에는 2000년 4월 현재, 513개의 데이터베이스가 등록되어 있다(표 1). DBcat의 주요 필드는 데이터베이스 이름, 설명, 도메인, 저작자, 웹주소 등을 가지고 있다.

3.2 미국립보건원

미국립보건원의 인간지놈연구소에서 유지하고 있는 분자 생물 데이터베이스 (http://www.oup.co.uk/nar

/Volume_28/Issue_01/html/gkd115_gml.html) 는 226개에 이른다. 이 데이터베이스는 분자 생물을 18개의 범주로 나누고 각 범주에 해당하는 데이터베이스명과 주소, 설명을 가지고 있다.

[표 1] DBcat에 등록되어 있는 DB 개수

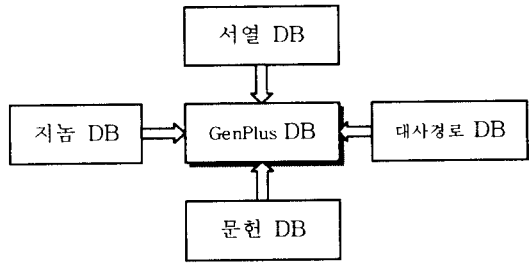
도메인	DB 개수
DNA	87
RNA	30
Protein	94
Genomic	58
Mapping	30
Protein structure	18
Literature	43
Miscellaneous	153
Total	513

[표 2] 미국립보건원에 등록되어 있는 DB 개수

도메인	DB 개수
Comparative Genomics	2
Gene Expression	12
Gene Identification and Structure	12
Genetic Maps	9
Genomic Databases	27
Intermolecular Interaction	3
Major Sequence Repositories	6
Metabolic Pathway and Cellular Regulation	10
Mutation Databases	32
Pathology	3
Protein Databases	36
Protein Sequence Motifs	12
Proteome Resources	4
RNA Sequences	21
Retrieval Systems and Database Structure	2
Structure	22
Transgenics	2
Varied Biomedical Content	11
Total	226

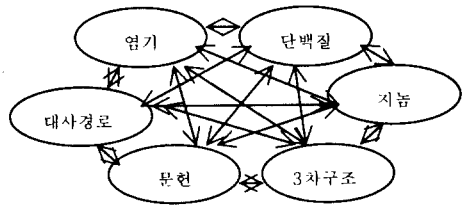
4. 생물정보 통합 검색 시스템

본 GenPlus데이터베이스는 서열 데이터베이스, 지놈 데이터베이스, 대사 경로 데이터베이스 그리고 문헌 데이터베이스를 연결한다(그림 1참조). GenPlus는 통합 환경에서 염기 및 단백질 서열, 생물의학 문헌 정보, 3차원 단백질 구조, 완전한 지놈을 검색할 수 있다(그림 2참조). 서열 데이터베이스는 염기 서열, 단백질 서열, 단백질 구조 서열 데이터베이스로 구성된다. 염기 서열 데이터베이스에는 GenBank, EMBL, DDBJ, PDB가 연결되고, 단백질 데이터베이스에는 GenBank, PDB, SwissProt, PIR이 연결되고, 단백질 3차 구조에는 PDB, MMDB가 연결되며 키워드, 단백질 서열, 염기 서열로 검색이 가능하다. 문헌 데이터베이스는 PubMed, Agricola, Brosis 가 연결되고, 대사 경로 데이터베이스에는 KEGG/Pathway가 연결된다.



[그림 1] GenPlus의 연결 구조

이질적인 서로 다른 데이터베이스에 연결하기 위해서 GenPlus는 1) 키워드; 2) Medline ID; 3) 접근ID를 사용한다. 각각의 분산된 데이터베이스에 자동으로 연결하기 위한 유일한 데이터베이스 정보를 유지한다. 사용자가 GenPlus에 질의하면 해당하는 데이터베이스에 접근하기 위해서 검색 엔진이 데이터베이스에서 매칭 엔트리를 검색한다(그림 3참조).

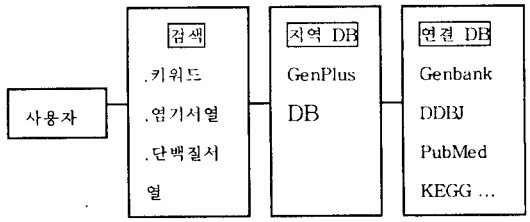


[그림 2] GenPlus의 통합 구조

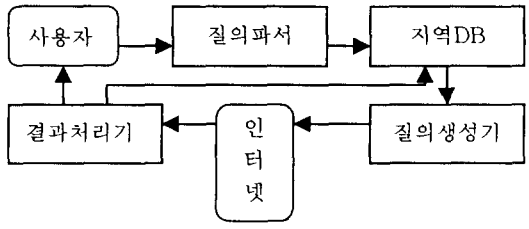
그림 2의 각 구성 요소에 연결되는 데이터베이스는 표 3과 같다.

[표 3] GenPlus에 연결 DB

DB 분류	연결 DB
서열 DB	염기-GenBank, EMBL, DDBJ, PDB 단백질-GenBank, PDB, SwissProt, PIR 3차 구조-PDB, MMDB
문헌 DB	PubMed, Agricola, Brosis
대사 경로	KEGG/Pathway
지놈 DB	GenBank



[그림 3] 질의 처리 구조



[그림 4] GenPlus구성도

그림 4에서 사용자가 질의를 입력하면 질의 파서는 입력된 정보를 검사하고 불필요한 정보를 제거한다. 지역DB에는 분산되어 있는 데이터베이스에 연결하기 위한 매칭 엔트리를 가지고 있다. 질의 생성기는 해당하는 분산된 데이터베이스에서 질의 처리가 가능하도록 프로파일을 재구성하고 웹 서버에 접속한다. 결과처리기는 각 웹 서버로부터 결과를 수집하고 수집된 결과를 분석하여 유용한 정보로 재구성하여 지역DB에 저장하고 사용자의 웹 브라우저나 전자우편으로 결과를 보낸다.

제안하는 시스템은 Mediline ID와 GenBank의 연결과 같은 사실 연결에서 시작하여, BLAST, FASTA의 결과에 의한 유사 연결, KEGG의 Pathway에서 대사경로를 통한 의미 연결을 지원한다. 제안한 시스템에서 키워드로 검색할 경우에 Medline ID, 서열 DB에 대한 접근 ID를 제공하고 표 2에 따른 분류에 따라서 관련 사이트 링크를 제공한다.

5. 결론

전 세계적으로 500여개의 생물정보 데이터베이스가 분산해 있다. 본 연구에서는 분산되어 있는 이질적인 생물정보 데이터베이스의 통합 검색 시스템을 제안하였다. 제안한 GenPlus에서는 염기 서열, 단백질 서열, 키워드 등을 이용해서 관련된 서열 정보, 단백질 구조 정보, 지놈 정보, 문헌 정보를 사용자에게 제공한다. GenPlus는 질의 파서, 매칭 엔트리를 저장하고 있는 지역 DB, 질의 생성기, 결과 처리기로 구성된다. GenPlus에는 전세계의 유명 생물정보 데이터베이스 12개가 연결된다.

참고문헌

[1] Andreas D. Baxevanis, "The Molecular Biology Database Collection: an online compilation of relevant database resources," Nucleic Acids Research, Vol. 28, No. 1, 2000.

[2] Minoru Kanehisa, "Trends guide to BIOINFORMATICS," pp.24-26, 1998. (http://bric.postech.ac.kr/topic/38.html)

[3] Claude Discala, Xavier Benigni, Emmanuel Barillot, and Guy Vaysseix, "Dbcat: a catalog of 500 biological databae," Nucleic Acids Research, Vol. 28, No. 1, 2000. (http://www.infobiogen.fr/service/dbcat)

[4] Arens, Y., Knoblock, C.A and Sheng, W.M, "Query Reformulation for Dynamic Information Integration," Journal of Intelligent Information Systems, Vol. 6 (2/3), pp. 99-130, 1996.

[5] Andleigh, P.K and Gretzinger, M.R., "Distributed Object-Oriented Data Systems Designs," PTR Prentice-Hall, Inc., Englewood Cliffs, 1992.

[6] Fran Lewitter, "Trends guide to BIOINFORMATICS," pp.3-5, 1999. (http://bric.postech.ac.kr/topic/39.html)

[7] Mousheng Xu, Susan Gauch, "Associated Biological Information Retrieval from Distributed Databases," ACM CIKM conf. pp. 193-200, 1998.

[8] Claudine Medigue, Francois Rechenmann, Antoine Danchin and Alan Viari, "Imagene: an integrated computer environment for sequence annotation and analysis," Vol. 15, no. 1, pp.2-15, 1999.

[9] 원세연, "생물정보 분야의 개괄 및 전망," 정보과학회지 제17권 제5호, pp.49-56, 1999.