Stand-Alone BLAST를 이용한 향상된 통합 서열분석시스템의 설계 및 구현

박춘구[0], 허정호, 최지인, 박윤주, 정동수, 남홍길
포항공과대학교 생물학전문연구정보센터
(madreach, hjh, jiin, oct1001, viroid97, hgn)@bric.postech.ac.kr

Design and Implementation of Advanced Sequence Analysis System using the Stand -Alone BLAST

Chungoo Park[0], Jeong Ho Huh, Ji In Choi, Yun Ju Park,
Dong Soo Jung, Hong Gil Nam
Biological Research Information Center, Pohang University of Science and Technology

요 약

오늘날 급속하게 발전하는 유전자 분석기술은 유전자 서열(sequence), 단백질의 기능(function) 및 구조(structure)정보와 같은 생명현상의 연구에 필수적인 정보들을 제공하게 되었다. 특히, 인간 유전체 프로젝트의 완성 이후 염기 및 단백질의 서열데이터를 이용하여 유사한 서열데이터의 검색 및 관련 단백질의 기능, 구조 정보들과 같은 생물정보의 종합적인 검색이 요구되고 있다. 하지만, 기존 대부분의 통합 서열분석시스템들은 단지 관련 정보를 포함하는 데이터 베이스들에 접근하여 서열유사성을 분석한 후, 그 결과를 단순히 디스플레이 하는 것이 대부분 이였다. 부연하면, 기존 통합 서열분석시스템들은 각 데이터베이스로부터 검색된 결과들 간의 명확한 관계를 설명하지 못하여 종합적인 생물정보를 제공하지 못하고 있다. 따라서 본 논문에서는 염기 및 단백질의 서열데이터로부터 서열유사성 검색 및 관련 단백질의 기능, 구조정보에 해당하는 종합적인 생물정보를 효과적으로 검색, 서비스 할 수 있는 통합 서열분석시스템의 설계, 구현에 관해 기술한다.

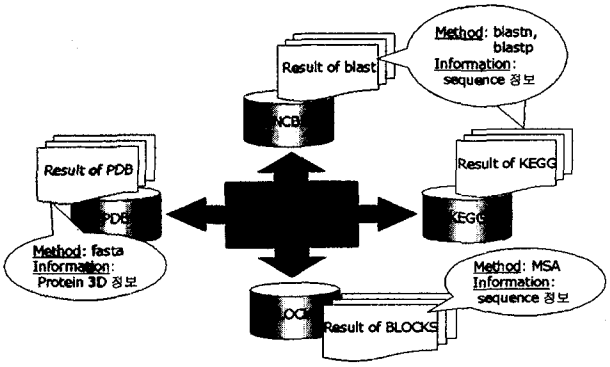
1. 서론

최근, 인간 유전체 프로젝트(HGP: Human Genome Project)의 완성과 분자 생물학의 발전으로 대용량(High-Throughput)의 gene sequence, EST, RNA structure, Protein function & structure, metabolic pathway, protein-protein interaction등의 다양한 생물학 정보들이 쏟아져 나오고 있다. 특히, 염기(DNA)와 단백질(Protein)의 서열정보 그리고 관련 단백질의 기능 및 다차원 구조정보는 생명현상의 연구에 밀접한 관계가 있어 그 중요성이 더욱 크다. 이러한 생물정보들은 각 데이터의 특징에 따라 적절히 구분되어 데이터베이스로 구축되어 있다. 대표적인 데이터베이스의 예를 들면, 염기(DNA) 및 단백질(Protein)의 서열정보를 저장하고 있는 GenBank[1], Swiss-Prot[2], TrEMBL[2], 단백질의 기능정보를 포함하고 있는 PIR[3], 단백질의 구조정보를 중심으로 구축되어진 PDB[4]가 있다[5].

염기 및 단백질의 서열분석을 통하여 관련 단백질의 기능 및 구조정보를 검색하고자 하는 생물학자들은 GenBank, Swiss-Prot, PIR, PDB 데이터베이스에 각각 서열유사성 검색을 수행하여 관련 정보를 검색하게 된다. 이러한 작업을 효과적으로 수행하기 위한 통합 서열검색 시스템[6]이 존재하긴 하지만, 이러한 검색방법은 각각의 데이터베이스가 동일한 서열유사성 검색 알고리즘(blast[7])을 이용하기 때문에 중복된 검색을 통한 계산자원(*computation resource*) 의 낭비를 초래하고, 또한 각 데이터베이스에 의해 검색된 결과들 사이에 관계를 명확하게 살펴보기 어렵기 때문에 염기 및 단백질의 서열정보에서부터 해당 서열의 기능 및 구조정보까지의 통합정보를 살펴보기 힘들다.

따라서, 본 논문에서는 계산자원의 낭비를 최소화 함과 동시에 염기 및 단백질의 서열분석정보에서부터 단백질의 기능 및 구조정보에 해당하는 종합적인 생물정보를 검색할 수 있는 통합 서열검색시스템을 설계, 구현한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 생물학정보센터[8]에서 개발한 통합 서열검색시스템 GeneNet[6]에 대해 살펴보고, 3장에서는 향상된 통합 서열검색시스템의 기본 개념과 설계 및 구현에 관한 전반적인 사항에 대하여 설명한다. 마지막으로 4 장에서는 결론 및 향후연구에 대하여 기술한다.



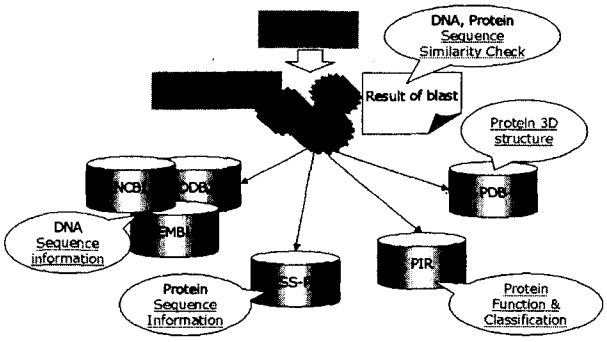
[그림 1] GeneNet의 기본 동작 원리

2. 관련연구

GeneNet[7]은 인터넷 상에 존재하는 다양한 생물학 데이터베이스를 이용하여 염기 및 단백질의 서열분석을 보다 쉽고 효과적으로 수행하기 위한 통합 서열분석시스템이다. [그림 1]과 같이 GeneNet은 사용자가 요청한 입력 서열을 4 개의 데이터베이스(GenBank, BLOCK, PDB, KEGG)에 각각 서열유사성분석을 요청한다. GeneNet은 검색된 서열분석 결과를 합하여 사용자에게 보여준다.

GeneNet은 정확한 서열분석을 수행하기 위해 4개의 데이터베이스에 반복적인 서열유사성분석을 수행하여 결과를 얻는다. 하지만, 대부분의 생물학 데이터 베이스들은 관련 정보들의 공유를 위하여 상호참조(cross-reference) 하고 있다. 특히, NCBI 는 EMBL, DDBJ, Swiss-Prot, PDB, PIR 등의 다양한 데이터베이스를 공유하고 있다[9]. 따라서, NCBI의 데이터를 이용한 서열분석(blast[7])결과는 BLOCK을 제외한 나머지 2개의 데이터베이스(PDB, KEGG)의 서열분석결과와 중복되게 되어 계산자원의 낭비를 초래한다. 또한, 4 개의 데이터베이스가 각자의 데이터 출력형태를 가지고 있기 때문에 NCBI 의 Blast 검색결과에 익숙한 대부분의 생물학자들은 모든 검색 결과를 정확하게 분석하기 어렵고, 그들 사이의 관계를 살펴보기가 어렵다. 예를 들어, NCBI의 GenBank를 이용하여 서열유사성 분석을 한 후 그 결과 중 특정 서열의 metabolic pathway와 단백질의 3차원 구조정보를 KEGG와 PDB의 검색결과를 통하여 명확하게 살펴보기 힘들다. 왜냐하면, PDB와 KEGG는 각각 PDB ID와 효소(enzyme) ID를 이용하여 단백질의 3 차구조와 단백질의 metabolic pathway정보를 검색할 수 있기 때문이다.

다음 장에서는 이와 같은 단점을 해결하기 위한 방안과 이를 적용한 향상된 통합 서열분석시스템에 관한 전반적인 사항에 대하여 기술한다. 참고로, 위에서 예외로 언급된 BLOCK의 경우는 마지막 장의 결론 및 향후연구에서 보다 자세히 기술한다.



[그림 2] 향상된 통합 서열분석시스템의 정보 흐름

3. 향상된 통합 서열분석시스템

3.1 기본 개념

기존 통합 서열분석시스템(GenNet)의 문제점을 정리하면 다음과 같다.

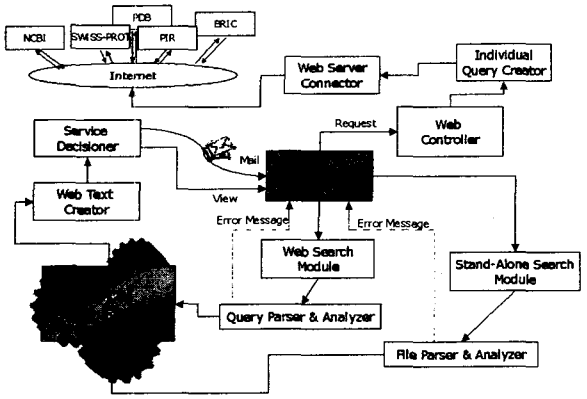
- 다양한 데이터베이스(NCBI, PDB, BLOCK, KEGG)의 검색 결과들 사이의 관계를 이해하기 어렵다.
- 염기 및 단백질의 서열정보를 이용하여 서열 유사성, 단백질 기능 및 구조 정보까지의 총괄적인 유전정보를 파악하기 어렵다.

따라서 향상된 통합 서열분석시스템은 다음과 같은 문제점을 해결하기 위하여 기존의 검색방법과는 다르게 NCBI의 Blast 검색결과를 중심으로 유전정보를 분석한다. NCBI는 EMBL, DDBJ, Swiss-Prot, PDB, PIR 등의 다양한 데이터베이스와 공유하기 때문에 NCBI의 Blast 검색결과는 관련데이터베이스의 서열유사성 분석결과를 포함할 수 있다. 또한 대부분의 생물학자에게 익숙하기 때문에 기존의 다양한 데이터베이스의 검색결과 분석의 어려움을 해결할 수 있다. 그리고 Blast 서열검색결과로부터 단백질의 기능 및 구조정보를 검색할 수 있는 index(PIR의 ID와 PDB의 ID)를 추출할 수 있다.

향상된 통합 서열분석시스템은 이와 같은 방법을 통해 최소한의 계산자원을 이용하여 서열분석정보에서부터 단백질의 기능 및 구조정보에 해당하는 총괄적인 유전정보를 파악할 수 있다[그림 2]. 참고로 [표 1]은 NCBI의 Blast 검색결과로부터 식별할 수 있는 데이터 베이스들의 리스트를 보여준다.

[표 1] Sequence Identifier Syntax

Database Name	Identifier Syntax
GenBank	gb|accession|locus
EMBL Data Library	emb|accession|locus
DDBJ	dbj|accession|locus
NBRF PIR	pir||entry
Protein Research Foundation	prf||name
SWISS-PROT	sp|accession|entry name
Brookhaven Protein Data Bank	pdb|entry|chain
Patents	pat|country|number
GenInfo Backbone ID	bbs|number
General database identifier	gnl|database|identifier
NCBI Reference Sequence	ref|accession|locus



[그림 3] 향상된 통합 서열분석시스템의 구조

추가적으로 [그림 2]에서 살펴볼 수 있는 바와 같이, 향상된 통합 서열분석시스템은 대용량의 서열분석을 효과적으로 수행하기 위하여 NCBI에서 제공하는 Stand-Alone Blast 엔진을 이용하여 Local에서 서열유사성분석을 수행한다.

향상된 통합 서열분석시스템의 정보흐름을 정리하면 다음과 같다.

(1)염기 및 단백질 서열입력 ➔ (2)Local Blast 엔진을 이용하여 서열유사성 분석 ➔ (3)서열분석결과로부터 각 데이터베이스(NCBI, SWISS-PROT, PIR, PDB)를 검색할 수 있는 키(ID)값을 파싱 ➔(4)검색 가능한 데이터베이스로부터 유전정보를 통합 출력

3.2 동작원리

[그림 3]은 향상된 통합 서열분석시스템의 전체적인 동작원리를 보여주고 있다. 향상된 통합 서열분석시스템은 사용자(End User) 로부터 서열정보를 입력 받아(Web Search Module) 파싱(Parsing) 및 에러처리를 수행한다(Query Parsing & Analyzer). 파싱된 정보들은 시스템내의 Blast 엔진에 넘겨지고, Blast 엔진을 통해 검색된 결과와 관련 추가정보를 HTML문서 형태로 만들어(Web Text Creator) 사용자에게 보여준다. 사용자는 HTML 문서를 통해 Internet상의 관련 데이터베이스에 접근하여 추가정보를 검색 할 수 있다(Web Controller, Individual Query Creator, Web Server Connector).

참고로, 본 시스템은 대용량 서열정보의 분석처리를 쉽게 하기 위하여 서열정보를 파일형태로 입력 받을 수 있게 했다(Stand-Alone Search Module). 이때 입력 받는 파일형태는 FASTA[10] 파일 형태이다.

3.3 구현환경

향상된 통합 서열분석시스템의 구현 환경은 다음과 같다.

- 운영체제: Linux 2.2.17
- 프로그래밍 환경: apache 1.3.22, python 2.2.1
- Blast 엔진: Blast version 2.0
- 사용된 서열DB종류: nr(단백질), nt(염기)

4 결론 및 향후 연구

본 논문에서는 Local Stand-Alone Blast 엔진을 이용하여 염기 및 단백질의 서열데이터로부터 서열유사성 검색 및 관련 단백질의 기능, 구조정보에 해당하는 종합적인 생물정보를 효과적으로 검색, 서비스 할 수 있는 통합 서열분석시스템에 관해 살펴보았다. 본 시스템은 염기 및 단백질 서열분석 관련 데이터베이스와 분석 소프트웨어에 익숙하지 않는 생물학자들에게 유용할 것으로 기대된다.

향후 본 시스템은 클러스터링 기반 병렬 Blast 엔진과의 연동을 통해 생물학정보센터[8]에서 서비스 할 예정이다. 그리고 2 장에서 언급했던 바와 같이 BLOCK 은 NCBI 의 데이터베이스와는 그 접근 방식이 다르기 때문에[11] 지금의 서열분석시스템에는 함께 추가할 수 없다. 따라서 향후 BLOCK 와 같은 profile 기반 데이터 베이스들(PFAM, PRINTS, PROSITE, ProDom, SMART) 을 위한 서열분석 모듈을 추가할 예정이다.

[참고문헌]

[1] http://www.ncbi.nlm.nih.gov/Genbank/index.html

[2] http://www.expasy.ch/sprot/

[3] http://pir.georgetown.edu/pirwww/

[4] http://www.rcsb.org/pdb/index.html

[5] Andreas D. Baxevanis, The Molecular Biology Database Collection: 2002 update, Nucleic Acids Res, 30:1-12. 2002.

[6] Dong-Sun Park . et al, GeneNet: A Meta - Sequence Similarity Analysi s Currents in Computational Molecular Biology. Universal Academy Press, Inc., TOKYO, Japan, 2000.

[7] Altschul, Stephen F. et al , "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389 -3402, 1997.

[8] http://bric.postech.ac.kr/

[9] ftp://ftp.ncbi.nih.gov/blast/documents/blastdb.txt

[10] http://www.ncbi.nlm.nih.gov/BLAST/fasta.html

[11] http://www.blocks.fhcrc.org/