3D 단백질 구조 데이터베이스 및 유사성 검색 시스템 구축†

이영화, 박성희, 류근호
충북대학교 데이터베이스 연구실
{lrh, shpark, khryu}@dblab.chungbuk.ac.kr

Building of Protein 3-D Structure Database and Similarity Search System

Rong-hua Li, Sung-Hee Park, Keun Ho Ryu
Database Laboratory, Chungbuk National University

요약

단백질 3차 구조 정보는 PDB에서 플랫화일 형태로 제공되고 있으며 이러한 플랫화일 각각의 엔트리들은 단백질 3차 분자 구조를 구성하는 원자들의 공간좌표정보, 서열정보, 실험정보 및 참조정보 등으로 구성된다. 이러한 정보들을 포함하고 있는 플랫파일로부터 필수적인 구조정보 및 서열정보 등의 효율적 검색을 위해서는 플랫파일을 데이터베이스로 구축함과 동시에, 구축된 데이터베이스를 위한 유사성 검색시스템 구축이 요구된다.

따라서, 이 논문에서는 Protein DataBank에서 제공하는 플랫파일을 공간객체 모델링기법에 기반한 관계형 데이터베이스로 구축하고 PSI-BLAST를 적용하여 단백질 서열 유사성 검색 시스템을 구축한다. 이렇게 함으로써 단백질 3차 구조 분자를 구성하는 원자에 대한 검색과 구조에 대한 서열 유사성 검색을 통하여 단백질 3차 구조 분류 및 구조 예측 시스템 구축에 활용할 수 있다.

1. 서론

HGP이후 국내외에서 생물정보학분야에서 유전체의 기능을 밝히기 위한 기능 유전체학이 활발히 연구되고 있다. 국내에서도 단백질 기능분석의 연구를 위해서는 다음과 같은 기술이 요구된다. 첫째, 단백질 서열 및 구조정보의 활용을 위해 이러한 데이터 저장을 위한 데이터베이스를 구축해야 한다. 둘째, 서열 및 구조데이터베이스를 기반으로 상동성 분석을 위한 유사성 검색 시스템[1,2] 구축이 필요 된다. 셋째, 장기적으로는 생물데이터 마이닝을 통해 단백질 구조를 분류하고 새로운 패턴을 추출하여 구조를 모르는 단백질 구조 및 기능 예측을 위한 시스템이 구축 기술이 필요하다.

기존의 유사성 검색 시스템은 BLAST[1,2]나 FASTA[9,10] 알고리즘을 이용하여 Genpept, PIR, SWISS-Prot과 같은 nr(non-redundent) 데이터 베이스에 구축되어졌다. 특히 NCBI의 BLAST검색 시스템[12]이 대표적이다. 그러나 이러한 유사성 검색 시스템은 단백질 서열 데이터베이스에 대해서 유사성 검색을 수행하기 때문에 단백질 구조의 유전적 변이 정보를 분석하기에는 불충분하다[11]. 따라서, 단백질 구조 분석을 위해서는 단백질 구조 정보를 포함한 서열 데이터 베이스에 대한 유사성 검색 시스템이 필요하다.

이 논문에서는 단백질 구조 정보를 포함한 PDB에서 제공하는 플랫파일을 공간객체 모델링을 적용하여 관계형 데이터베이스로 구축하고, BLAST의 결과를 다중 alignment로 생성하기 위해 Position-Specific score matrix를 사용하는 PSI-BLAST를 적용하여 단백질 서열 유사성 검색 시스템을 구축한다. 이 시스템은 PDB 서열에 대한 유사성 검색과 다른 단백질 서열 데이터 베이스의 유사성 검색 결과를 비교하여 구조의 유전적 변이를 분석하는데 활용할 수 있다.

2. 관련 연구

단백질 3차 구조 데이터베이스로 미국의 PDB (protein data bank)가 있다[3]. PDB는 생물학적인 단백질 3차원 고분자 결정 구조를 위한 데이터베이

† 이 연구는 2001년도 KISTI 위탁 연구비 지원으로 수행되었음

스로서 1971년 부룩헤이번 국립 연구소(BNL)에 의해 공개되었고, 80년대에 들어서 엑스레이 결정 구조 측정 (X-ray crystal structure determination), 세포 자기 잔향(Nuclear Magnetic Resonance-NMR), 저온 전자 현미경(Cryoelectron microscopy) 등의 방법을 이용하여 단백질 고분자 결정 구조를 찾아냈다. 2001년을 기준으로 모두 약 15000개의 단백질 엔트리를 보유하고 있다.

유사성 검색 프로그램에는 NCBI의 BLAST(Basic Local Alignment Search Tool)[1,2]와 와싱톤 대학에서 개발한 WU-BLAST[1,2,4,5]가 있다. NCBI의 BLAST는 데이터베이스 서열에서 워드 또는 k-tuple 검색을 통해 서열 alignment속도를 높이며, W, T, X와 같은 다양한 파라미터들에 대해 할당된 값들을 사용자가 임의적으로 지정하여 검색 민감도를 조절할 수 있고, 유사성 발견 가능성을 증가시키는 지역 alignment를 생성하며, 유사성 점수를 평가하는 메소드를 포함하여 중요도가 없는 매치에 대한 번역 가능성을 줄인다. WU-BLAST는 단백질과 염기서열을 표현하기 위한 XDF 포맷을 이용하여 데이터베이스 파일의 크기에 제한 없이 단백질 유사성 검색을 지원한다. BLAST는 검색속도와 검색 결과의 정확성이 데이터베이스 파일 크기의 영향을 받는 단점이 있다.

3. PDB 플랫파일의 특성 및 구조 분석

PDB의 3차원 단백질 구조에 대한 플랫파일은 3차원과 2차원 그리고 1차원 서열에 대한 정보뿐만 아니라 3차 구조에 대한 물리화학적인 주석과 기능적인 주석정보를 포함한다. PDB의 플랫파일은 주요하게 Title, Primary structure, Heterogen, Secondary structure, Connectivity Annotation, Miscellaneous feature와 Crystallographic & Coordinate transformation부분으로 구성된다. 이러한 플랫파일 레코드가 포함하는 정보는 표 1과 같다.

표 1 플랫파일 단백질 구조 정보

HEADER	Idcode를 포함한 PDB 엔트리 고유 식별자
TITLE	엔트리에 대한 실험과 분석에 대한 제목
COMPND	한 엔트리 고분자에 대한 설명
SOURCE	각각의 엔트리 분자에 대한 화학적, 생물학적 소스
DBREF	단백질 서열에 대한 참조 링크와 참조할 수 있는 서열 데이터베이스 엔트리
SEQRES	고분자 각 체인에서 아미노산 또는 핵산의 서열
HELIX	분자내에서 시작하고 끝나는 residue 위치, 이름과 Helix의 길이 정보
SHEET	분자내에서 시작하고 끝나는 residue의 위치와 이름
CRYST1	unit cell parameters, space group과 Zvalue 표현 crystallographic 실험에 대한 기하와 좌표 체계 변환값
ORIGXn	엔트리의 직교좌표를 나타내기 위한 원래의 좌표 변환값
ATOM	표준 residue에 대한 정보 및 이에 포함된 원자의 좌표값

4. 단백질 데이터베이스 설계

4.1 공간 데이터 모델을 적용한 단백질 구조 모델링

단백질 구조 모델링은 공간객체와 유사한 단백질 구조의 기하적 특성에 의하여 원자에 해당하는 노드와 각각의 원자들에 대한 연결관계를 라인으로 나타내는 네트워크 공간 데이터 모델링 기법을 적용한다.

4.2 단백질 3차 구조 스키마 설계

3절의 플랫파일 분석 결과를 이용하고 네트워크 공간 데이터 모델링 기법을 적용하여 단백질 서열 및 구조정보와 그 관계를 그림 1의 E-R다이어그램으로 표현하였다.

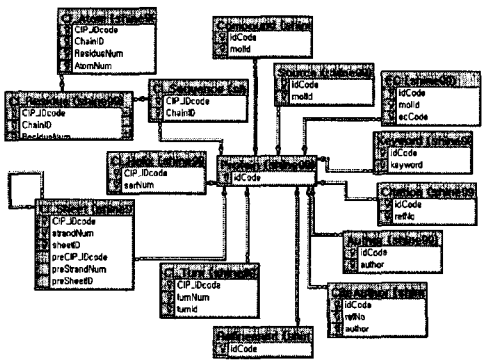

그림 1 단백질 구조 데이터 베이스의
E-R 다이어그램

5. 유사성 검색을 위한 PSI-BLAST 시스템

단백질 패밀리 발견을 위하여 BLAST2.0 버전에서 제공한 PSI-BLAST (Position-Specific Iterated BLAST) [6,7,8]는 단백질의 유사성 발견에 있어서 프로파일을 사용하여 데이터베이스 검색을 수행한다. 따라서, PSI-BLAST는 BLAST 검색을 통하여 생성된 결과 메트릭스를 입력 데이터로 사용하는 알고리즘을 채택하였고, 다중 alignment와 중요한 지역 alignment로부터 프로파일을 생성한다. 이러한

과정을 임의적으로 정해진 수만큼 반복 수행하여 유사성 검색을 수행한다.

PSI-BLAST가 유사성 검색을 수행하는 과정은 그림 2와 같다. 첫 번째 단계로, 하나의 단백질 서열을 입력으로 받고 gapped BLAST를 사용하여 단백질 데이터베이스와 비교한다. 두 번째 단계는 다중 alignment 생성에 따른 중요한 지역 alignment로부터 프로파일을 생성한다. 다음 단계로, 프로파일은 단백질 데이터베이스에 대해 비교되어지고, 네 번째 단계로 지역 alignment에 대해 통계적 중요성을 평가한다. 프로파일 substitution 점수들이 고정된 크기를 가지고 있고 gap에 대한 점수가 포지션에 대해 독립적으로 유지되기 때문에 gapped BLAST alignment를 위한 통계적 이론과 파라미터를 프로파일 alignment에 적용할 수 있다. 마지막 단계로 PSI-BLAST는 2번 단계로 수행을 반복하여 보다 뛰어난 유사성 검색을 수행한다.

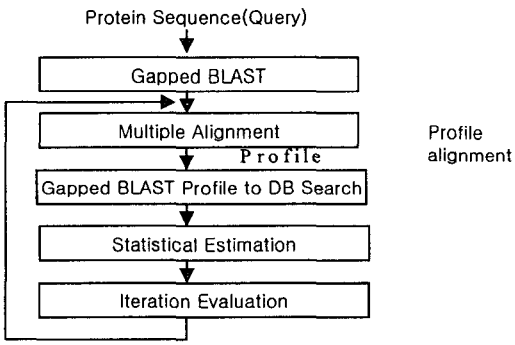

그림 2 PSI-BLAST 수행 5단계

6. PDB 구축 및 유사성 검색 시스템 구현

단백질 3차 구조 데이터베이스 구축과 PSI-BLAST의 적용에 대한 시스템으로 SunSPARC Ultra-250, Oracle7 및 PRO*C 언어를 사용하여 구축하였다.

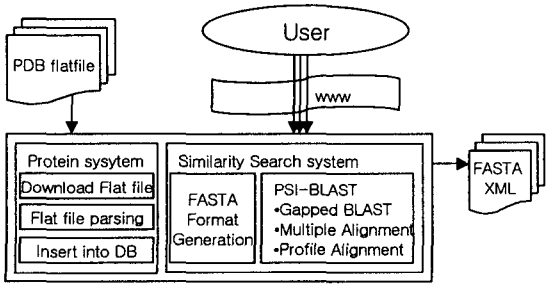

그림 3 단백질 구조 정보 유사성 검색 시스템

그림 3은 웹 상에서 제공하는 플랫파일을 관계형 데이터베이스로 재구성하여 구축하고, 이 데이터베이스를 이용하여 PSI-BLAST에서 사용할 수 있는 FASTA 포맷을 생성하는 것에 대한 구조를 보여주고 있다.

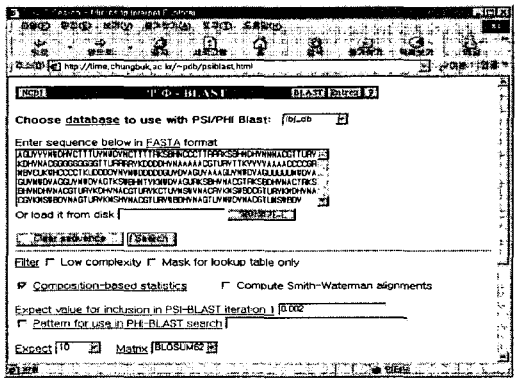

그림 4 FASTA 포맷을 이용한
단백질 유사성 검색 질의

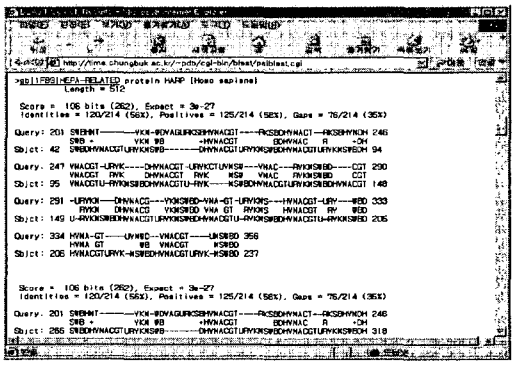

그림 5 PSI-BLAST를 이용한
단백질 서열 유사성 검색 결과

그림 4는 입력 FASTA 포맷의 서열에 대하여 단백질 구조 데이터베이스에 저장된 유사한 서열을 검색하기 위한 질의 창이며, 그림 5는 이러한 질의에 대하여 PSI-BLAST 검색을 수행한 결과 창이다.

7. 결론

이 논문은 PDB에서 제공하는 플랫파일을 분석하여, 단백질 관련 정보와 유사성 검색에 필요한 정보만을 추출, 단백질 3차 구조 정보 데이터베이스를 구축하였고, 이러한 단백질 3차 구조 정보 데이터베

이스를 이용하여 FASTA 포맷의 데이터를 생성 및 PSI-BLAST 검색 시스템에 적용하여 단백질 서열 유사성 검색을 수행하였다.

그러므로 단백질 3차 구조 분자를 구성하는 원자에 대한 정보들을 효율적으로 검색할 수 있으며, 단백질 3차 구조 분류를 위한 단백질 데이터들 간의 서열 유사성 및 상동성 검색 등에 활용하며, 단백질 3차 구조 분류 및 예측 시스템 구축에 적용할 수 있다.

향후 연구로는 이러한 단백질 3차 구조 정보 데이터 베이스를 이용하여 단백질 패밀리에 대한 분류 및 예측 시스템에 적용한다.

참고문헌

[1] Altschul, S. F., Carrol, R. J., and Lipman, D. J.(1990). Basic local alignment search tool. J. Mol. Biol., 215, 403.

[2] Karlin, S, SF Altschul., "Applications and statistics for multiple high-scoring segments in molecular sequences", Proc. Natl. Acad. Sci., 1993.

[3] Helen M. Berman, John WestBrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov and Philip E. Bourne, "The Protein Data Bank", Oxford Unversity Press, 2000.

[4] Karlin, S. & Altschul, S.F, "Methods for assessing the statistical significance of molecular sequence features by using general scorling schemes," Proc. Natl. Acad. Sci. USA 87, 1990.

[5] Altschul, SF, and W Gish. Local alignment statistics. ed. R R. Doolittle. Methods in Enzymology 266:460-80, 1996.

[6] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.D. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acid Res.1999

[7] Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.L., Koonin, E.V. & Altschul, L.F. "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," Nucleic Acids Res. 2001.

[8] David W. Mount, "Bioinformatics : Sequence and Genome Analysis" Cold Spring Harbor Laboratory Press, 2001.

[9] Lipman D.J., Pearson W.R., "Rapid Similarity Searches", Science vol. 227 pp1435-1441, 1985

[10] Wilbur W.J., Lipman D.J., "Rapid Similarity Searches of nucleic acid and protein data banks", 1983

[11] David W.Mount "Bioinformatics: Sequence and Genome Analysis" , 2001

[12] http://www.ncbi.nlm.nih.gov