공간객체 모델 기반 단백질 3차 구조 모델링

한욱[0] 박성희 이순희 류근호
충북대학교 데이터베이스 연구실
{yustuk[0], shpark, shlee, khryu}@dbalb.chungbuk.ac.kr

Modelling of three Dimensional Structure in Protein based on Spatial Object Model

Yu Han[0] Seng-Hee Park Sun-Hee Lee Keun Ho Ryu
Database Laboratory, Chungbuk National University

요 약

PDB에서 제공하는 단백질 3차원 고분자결정 구조에 대한 플랫파일은 원자들의 좌표, 서열정보, 실험정보 및 참조 정보가 포함된다. 이러한 정보를 포함하고 있는 플랫파일로부터 필수적인 구조정보 및 서열정보 등의 효율적인 검색을 위해서는 이러한 데이터를 추출하여 데이터베이스 구축이 요구되며 이 때 단백질 구조 및 서열 정보와 실험 및 참조 정보의 관계에 대한 모델링이 중요하다. 따라서 이 논문에서는 PDB에서 제공하는 플랫파일들의 엔트리들을 분석하고 3차원 공간 객체의 기하적 특성을 갖는 단백질 3차 구조를 공간객체로 표현하고 공간객체 모델을 적용하여 모델링한다. 이렇게 함으로써 단백질 3차 구조 분자를 구성하는 원자 및 구조 정보 검색이 가능하며 위상 및 기하 연산자를 이용하여 단백질 구조 분석에 활용할 수 있다.

1. 서 론

단백질은 각기 고유의 기능과 역할을 가지고 있지만 그 특성은 단백질을 구성하는 아미노산의 종류와 그들 각자의 배열순서에 의해 형성되는 입체구조에 따라 결정된다. 따라서 한 단백질의 구조 및 서열을 알면 그와 유사한 구조를 가진 단백질의 기능을 예측할 수 있다. 그 동안 단백질의 3차원구조와 서열을 해석하여 단백질의 기능적 단서를 찾기 위해 수많은 유사성 및 상동성 검색이 진행되어 왔다.

단백질 데이터는 그 양과 데이터의 범위가 상당히 넓고 크기 때문에 그 각각의 데이터 타입과 값들을 유연하게 처리해야 한다. 단백질 데이터에서 3차 구조 정보는 수많은 원자의 공간 좌표정보를 포함하고 있다.

단백질에 대한 데이터들의 저장소인 PDB(Protein Data Bank)[1,2]는 3차원 단백질 및 관련 데이터를 플랫파일 형태로 배포하고 있다. 이러한 플랫파일의 단백질 공간 좌표정보, 서열정보 등으로 구조데이터의 사용과 분석을 촉진하기 위하여 이런 데이터를 먼저 추출하여야 한다. 이어서 단백질의 3차 구조를 보다 직관적으로 표현하기 위하여 공간객체 모델링 방법을 적용한 데이터의 모델링이 선행되여야 한다. 따라서 공간객체 모델링을 적용한 보다 효율적인 새로운 데이터베이스 시스템의 구축이 필요하다.

이 논문에서는 공간 객체의 위상과 기하적 특성을 가지는 단백질의 구조적 특성을 분석하고 PDB에서 제공하는 플랫파일을 분석하여 공간객체 모델링 방법을 적용하여 단백질 3차 구조를 모델링한다. 이러한 모델링 기법을 적용하여 데이터베이스를 구축하면 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.

이 연구는 2001년도 한국과학기술정보연구원 Bioinformatics연구팀의 위탁연구비 지원으로 수행되었음

2. 관련연구

2.1 단백질 데이터베이스

이질적이고 매우 방대한 생물학적데이터들에 대한 통합 데이터베이스 검색 사이트 중에서 단백질 서열과 구조 정보를 보유하고 있는 사이트로 SWISS-PROT[1,3], PDB[1,2], FSSP[4,5], HSSP[6], MMDB[7]등이 있다.

2.2 공간객체 모델링

공간 객체를 관리하기 위한 공간 데이터베이스[9,10,11]는 공간 객체뿐만아니라 CAD, VLSI, 로봇공학과 영상 처리데이터에 기초한 응용프로그램 분야에서 매우 유용하게 사용된다.

공간적 특성을 나타내는 공간 객체는 (point)점, 선(line), 면(region)의 기본타입으로 기하적으로 표현되고 또한 복잡한 공간 객체 콜렉션은 스파게티, 네트워크 또는 위상적 모델로 표현한다.

공간객체 모델링은 공간패턴과 흐름 등의 분석에 있어서 다차원의 공간 정보를 동시에 분석 및 표현하고 예측과 설명을 위한 경우 정밀한 수치적 실험을 할 수 있고 다양한 그래

픽으로 표현할 수 있는 장점을 갖는다.

3. 단백질 3차 구조의 기하적 특성

3.1 단백질 3차 구조

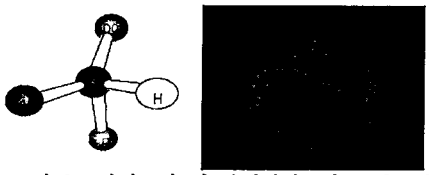

그림 3.1 아미노산 및 단백질의 3차구조[8,12]

대부분 단백질은 하나의 폴리펩타이드로 되어 있으며 이것이 복잡해지거나 굴곡을 이루면서 입체적 고차 구조를 형성한다. 단백질의 기능은 아미노산의 배열순서에 따라 결정되는 고차구조(입체구조)에 의해 나타난다. 단백질의 1차원 구조는 펩타이드 체인에서 아미노산들의 서열이다. 이러한 서열을 구성하는 아미노산 사이의 수소 결합을 통하여 2차원 구조를 형성한다. 2차원 구조는 아미노산을 구성하는 측쇄들 간의 상호작용에 의해 2차원 구조가 접히게(folding)되어 다음과 같은 3차원 구조를 형성한다.

3.2 공간 객체의 위상 및 기하

네트워크 공간 모델은 점과 폴리라인 사의의 위상관계가 저장될 수 있으며 그래프 기반 응용프로그램에서부터 시작되었다. 네트워크 모델에서는 점(point)과 면(polygon, region) 뿐만 아니라 노드(node)와 아크(arc)를 이용해서 모델링한다. 노드는 아크에 의해서 연결된 점으로써 아크의 끝점이거나 평면에서 고립된점이다. 일반적인 점은 다각형 또는 다른 선의 정점이 된다. 아크는 노드에서 시작해서 노드로 끝나는 폴리라인이다.

- point : *[x:real, y:real]*
- node : *[point, <arc>]*
- arc : *[node-start, node-end, <point>]*
- polygon :*<point>*
- region :*{polygon}*

그림 3.6은 네트워크 구조의 예이며 여기서 노드 n1은 아크 a1, a2, a3와 a4를 연결한다.

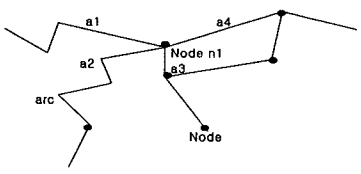

그림 3.6 공간 네트워크 모델 예[9]

3.3 단백질 3차 구조의 위상 및 기하

단백질 분자 구조에 포함된 각각의 원자는 3차원 공간좌표를 갖는다. 각 원자는 이를 식별할 수 있는 화학적 이름과 하나의 분자 내에서 원자를 식별할 수 있는 고유한 식별자를 갖는다. 단백질 분자는 하나 또는 그 이상의 하위 구조로 나누어진다. 하위 구조는 하위구조를 구성하는 원자들이 다른 것들에 비해 공간적으로 고정되어 회전할 수 없다면 하위그래프로 나타낼 수 있다. 이러한 하위구조는 전체 단백질 분자구조에서 회전할 수 있다. 하위 구조 내에 포함된 원자의 상대적 위치와 하위 구조 외부에서 원자는 회전에 의해서 변경될 수 있다. 따라서, 단백질 3차 분자 구조를 3D 네트워크 그래프로 고려하고 각각의 원자는 노드로 각 원자사이의 결합은 간선으로, 그리고 그래프의 블럭은 하위 구조로 표현한다. 두개의 하위 구조는 간선에 의해서 연결된다.

그림 3.7에서 숫자는 원자의 식별자이고 (a)처럼 괄호안의 문자는 원자의 이름을 나타낸다. 그리고 원자 {0,1,2,3,4}를 갖는 하위구조 S0와 {5,6,7,8,9}를 갖는 S1으로 구성된다. S0와 S1은 원자 4와 5의 결합으로 연결되며 이것은 노드4와 5에의해서 연결되는 arc={5, 6}로 표현할 수 있다.

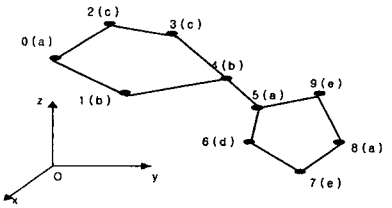

그림 3.7 단백질 3차 구조 공간 네트워크

4. 3차원 모델링을 위한 PDB 플랫파일 분석

PDB(protein Data Bank)의 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 feature 와 crystallographic & Coordinate transformation 부분으로 구성된다.

Title부분은 엔트리에 대한 실험과 생물학적 분자에 대한 설명을 포함한다. 주요한 레코드로는 HEADER, TITLE, COMPND, SOURCE, KEYWORD, EXPDTA, AUTHOR, JRNL 과 REMARK 이며 이 레코드들은 단백질 3차 구조에 대한 일반적인 내용과 분자에 대한 생물화학적 소스, 분자 구조를 발표한 저자 및 논문과 분자 검색 관련 정보이다. Primary structures부분은 DBREF와 SEQADV, SEQRES와 MODRES와 같은 분자를 구상하는 각각의 체인에 대한 residue 서열 정보와 서열의 참조 정보이다. Heterogen부분은 엔트리에 포함된 비표준 residue에 대한 설명을 포함한다. Secondary structure부분은 단백질의 2차 구조 요소인 HELIX, SHEET 와 TURNS와 폴리펩타이드 구조에 대한 정보를 제공한다. connectivity Annotation은

황화합물 결합 위치 및 존재와 다른 연결정보를 나타낸다. Miscellaneous feature부분은 분자의 활성사이트와 같은 분자의 특징을 설명한다. 다른 특징은 Title 부분의 remark레코드에서 설명된다. Crystallographic & Coordinate Transformation 부분은 결정학 실험의 기하와 좌표 체계시스템변환에 대한 정보를 포함한다. CRYST1 과 ORGXn 와 같은 주요 레코드를 포함하다. Coordinate 부분은 원자의 좌표에 대한 콜렉션 정보인 ATOM, MODEL과 ENDMODEL 레코드를 포함한다. ATOM은 단백질 3차 구조에 대한 좌표 정보 및 원자를 포함하는 residue에 관련 정보를 포함한다.

5. 단백질 3차 구조 모델링

단백질 3차 구조 모델링은 3.3에서 언급한 공간객체와 유사한 단백질 구조의 기하적 특성에 3.2에서 언급한 네트워크 공간 객체 모델링 방법을 적용하여 그림 5.2와 같이 모델링하

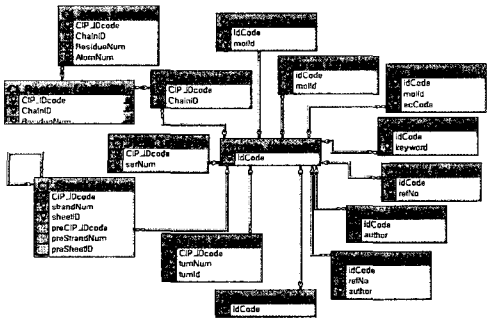

그림 5.2 데이터베이스 구축을 위한 ER 다이어그램

였다.

우선적으로 단백질 분자에 대한 일반적인 정보와 구조 정보를 서로 다른 테이블에 나누어 모델링한다. 구조정보의 모델링에서는 3차 구조 중 좌표 값을 포함하는 원자 정보는 노드로 표현되고 원자를 포함하는 residue는 하나의 하위 구조가 되며 이러한 residue의 결합이 2차구조요소인 α-Helix와 β-sheet을 나타내며 전체 하나의 분자구조를 나타낸다. 또한 분자의 1차구조적 측면에서는 residue가 노드로 표현되고 이들간의 연결을 아크로 표현하여 분자구조를 모델링한다.

6. 결론

HGP이후 기하급수적으로 증가한 공개용 생물 데이터를 국내외의 생명공학 연구에 활용하기 위해서 유전체 정보 데이터베이스 구축 및 분석을 위한 소프트웨어 개발이 활발히 진행되고 있다. 특히 국내 단백질연구 분야의 활발한 연구를 위해서는 단백질 서열 및 구조 데이터베이스의 구축 및 구축된 데이터베이스이 최신정보 유지가 필수적이다.

이 논문에서는 공간 객체의 네트워크 위상과 기하적 특성을 가지는 단백질의 구조적 특성과 PDB에서 제공하는 플랫 파일을 분석하였다. 이 분석 결과를 기반으로 feature-based 지리객체 모델링 방법과 네트워크 공간객체 모델링 방법을 적용하였고 단백질 3차 구조를 모델링하였다.

이 연구에서 제공한 모델링 기법을 적용하여 데이터베이스를 구축하면 단백질 3차 구조 분자를 구성하는 원자에 대한 정보들을 효율적으로 검색할 수 있다. 또한 단백질 3차 구조 분류를 위한 단백질 데이터들 간의 서열 유사성 및 상동성 검색 등에 활용할 수 있고, 단백질 3차 구조 분류 및 예측 시스템 구축에 적용할 수 있다.

향후 연구로는 이러한 단백질 3차 구조 데이터베이스에 대해서 geometry의 연산을 이용하여 유사성 검색 질의를 처리할 수 있도록 공간 연산자의 확장하는 것이다.

참고문헌

[1] H. M. Berman, J. W. Brook, ZukangFeng, Gary Gilliland, T. N. Bhat, *Helge Weissig, Ilya N. Shindyalov and Philip E.* Bourne, "The Protein Data Bank", Oxford University Press, 2000.

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

[3] D. Higgins, W. Taylor, "Bioinformatics : Sequence, Structure, and databanks", Oxford University press, 2000.

[4] L Holm and C Sander, "Dali/FSSP classfication of *three-dimensional protein folds*" *Nucleic Acids Research*, Oxford University Press, 1997.

[5] Liisa Holm and Sander, "The FSSP database: fold classification bassed on structure-structure alignment of proteins", Oxford University Press, Nucleic Acids Research, 1996.

[6] Chris Dodge, Reinhard Schneider, Chris Sander, "The HSSP database of protein structure-sequence alignment and family profiles", Oxford University Press, Nucleic Acids Research, 1998.

[7] Aron Marchler-Bauer, Kenneth J. Addess, Colombe Chappey, Lewis Geer, Thomas Mad더, Yo Matsuo, Yanli Wang and Stephen H. Bryant, "MMDB: Entrez's 3D structure database", Oxford University Press, Nucleic Acids Research, 1998.

[8] T.A. Brown, "Genomes", Bios Scientific publishers Ltd, 1999

[9] Philippe Rigaux, Michel School, Agnes Voisard, "Spatial Database", Morgen Kaufmann, 2001.

[10] OpenGIS Consortium, Inc. OpenGIS, Simple Fea-tures Specification For OLE/COM Revision 1.1, OpenGIS Project Document 99-050, 1999.

[11] Agatha Y. Tang, Teresa M. Adams, E. Lynn Usery, Spatial Data Model Design for Feature-Based Geographical Information Systems", Int. Jour. of Geographical Information Science vp.10 No.5,1996.

[12] 오까기치비, 아비꼬 요시미쯔," 생명과학을 위한 비주 얼생화학·분자 생물학", 해돋이, 1997.