단백질 3차 하위구조 비교 시스템 설계

유남희*, 정광수*, 손교용*, 정용제**, 류근호*
*충북대학교 데이터베이스/바이오인포메틱스연구실, **충북대 생명과학부
e-mail:{nami,ksjung,gysohn,khryu}@dblab.chungbuk.ac.kr,
chungyj@chungbuk.ac.kr

Designing of Comparison System for Protein Tertiary Substructure Database

Nam Hee Yu*, Kwang Su Jung*, Gyo Yong Sohn, Yong Je Chung**,
Keun Ho Ryu*
*Database/Bioinformatics Laboratory,
**Division of Life Science Chungbuk National University

요 약

생명체 내에서 기능 수행 시 각종 물질들이나 단백질들끼리 상호결합을 해야 한다. 이런 결합성을 결정짓는 것들이 단백질의 3차원 구조이기 때문에 단백질 구조연구는 중요하다. 이 논문에서는 단백질 구조데이터 및 관련된 구조정보의 통합된 데이터베이스를 구축하고 웹 환경에서 질의된 단백질과 유사성 비교를 진행하여 그 결과 및 연관된 정보를 검색하여 체계적으로 정보를 제공하는 단백질 구조 비교시스템을 제안한다.

제안 시스템을 구축하기 위하여 공개용 단백질 구조데이터 저장소인 Protein Data Bank의 플랫파일에서 필수적인 구조데이터정보만을 추출하여 여기에서 단백질의 하위구조 생성 알고리즘을 적용하여 데이터베이스를 구축한다. 사용자가 인터넷을 통하여 진행한 질의는 하위구조처리 모듈을 통하여 하위구조를 생성하고 구조유사부분에 대해 RMSD값이 계산되고 이와 연관된 구조정보의 검색이 진행 된 후 체계적으로 출력화면에 보여준다. 제안 시스템은 단백질의 전체적인 서열과 구조 정보를 이용하지 않고서, 단백질 기능을 결정하는 핵심영역을 포함하는 표면을 효과적으로 비교함으로써 기존의 구조비교 시스템보다 빠른 검색과 상세한 분석을 지원한다.

1. 서론

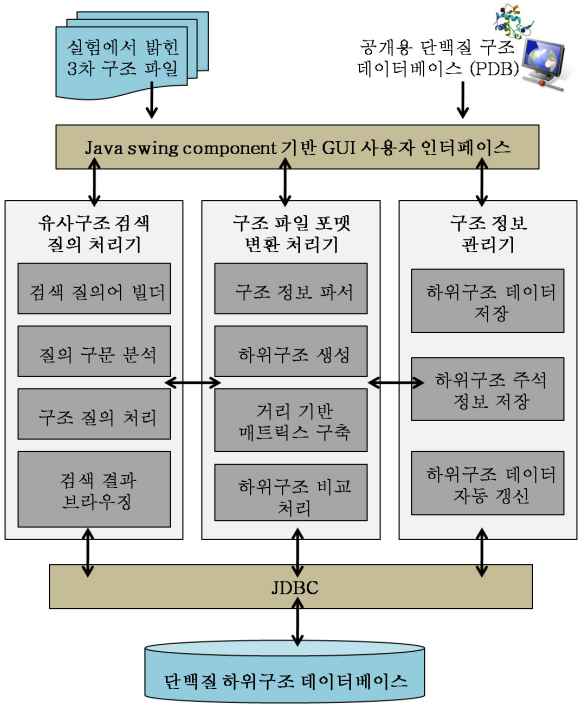
기능을 모르는 단백질 서열이 있을 때 그것과 관계가 있다고 생각되는 서열을 찾는 것은 생명과학 분야에서 중요한 첫 걸음이다. 종종 서열상에서의 유사성은 없지만 3차 구조상에서의 유사성을 가진 단백질들이 발견되곤 한다. 이는 단백질의 기능이 구조적으로 잘 보존된 특정한 영역에 따라 결정되기 때문이다.[1] 서열의 잘 보존된 부분의 유사성을 통해 단백질의 특정 부분의 기능을 예측할 수 있는 것처럼 구조의 부분 유사성 또한 특정 부분의 기능을 예측할 수 있다. 이 논문에서는 단백질의 기능을 예측할 수 있는 구조의 부분 유사성 비교시스템을 제안한다.

단백질 3차 구조 파일인 PDB[2] 플랫파일은 구조 데이터뿐만 아니라 모든 원자의 위치정보를 가지고 있다. 이 논문의 하위구조 생성 알고리즘[3][4][5]은 C_α원자와 C_β원자의 좌표정보만을 이용하여 계산하는 방법을 이용하기 때문에 단백질의 하위구조비교에 필요한 데이터들만 추출한다. 추출된 정보는 단백질 하위구조 처리 모듈을 통하여 거리기반 매트릭스를 생성하여 하위구조비교 알고리즘에 필요한 플랫파일 포맷으로 저장한다. 새로운 플랫파일 포맷은 하위구조비교에 필요한 단백질의 3차구조정보만 들어있기 때문에 데이터 파일을 읽는 속도를 많이 단축할 수 있다. 이 단백질과 연관된 구조주석 정보는 우리가 설계한 데이터베이스에 저장된다. 그리고 사용자가 웹을 통하여 전송한 단백질 구조 혹 단백질 식별자 코드로 우리가 제안한 하위구조 비교시스템[6]을 이용하여 하위구조 비교가 진행된다. 질의 단백질과 데이터베이스 내 단백질의 구조적을 매치되는 부분과 RMSD[7]값이 단백질의 주석정보와 함께 사용자에게 보여준다. 우리 시스템은 검색 속도가 빠르고 질의 단백질 구조와 연관된 단백질의 정보를 손쉽게 찾아낼 수 있다. 단백질 기능에 대한 구조 비교를 실험을 통해 밝혀내는 것은 실험해야할 대상이 많기 때문에 많은 시간과 비용이 소요된다. 하지만 우리 연구의 단백질 하위구조 비교시스템으로 정확한 단백질의 기능 후보들을 선별하기 낼 수 있다면 실험 시간과 비용을 획기적으로 축소시킬 수 있다. 또한 단백질은 생물체 내의 물질대사에 직접 작용하므로 연구 결과를 신약개발이나 생물의 생명현상을 밝히는데 응용할 수 있다.

"이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구(No. R11-2009-014-02002-0)이며 2009년 교육과학기술부로부터 지원받아 수행된 연구임"(지역거점연구단육성사업 / 충북BIT연구중심대학육성사업단)

2. 시스템 설계

그림 1에서 단백질 하위구조 비교시스템은 구조데이터 플랫파일, 단백질 유사구조 검색 질의 모듈, 단백질 하위구조 처리모듈, 단백질 하위구조 저장 관리 모듈 및 입출력 인터페이스로 구성되어있다. 하위구조 데이터베이스 구축은 새로 배포된 PDB 엔트리들을 주기적으로 다운로드 하여 구조데이터를 추출하고 단백질 하위구조 처리 모듈을 이용하여 새로운 플랫파일형식에 맞추어 구조데이터베이스를 구축한다. 다음 필요한 주석데이터는 주석정보데이터베이스에 입력된다. 따라서 입출력 인터페이스는 사용자의 질의를 받아서 하위구조비교와 주석데이터검색을 진행한 후 다시 웹을 통하여 보여 진다.



(그림 1) 전체 시스템 구성도

3. 하위구조 비교 시스템

3.1 단백질 유사구조 검색 질의 처리 모듈

단백질 유사구조 검색질의 처리모듈은 검색 질의어 빌더와 구문 분석을 통해 단백질 3차 구조 데이터를 추출하고 하위구조 처리 모듈에서 구조 질의를 처리하여 검색결과를 보여준다. 이 모듈의 하위 컴포넌트들은 다음과 같다. 검색 질의어 빌더는 단백질의 식별자 코드와 구조, 이름 등 과 AND와 OR 연산자를 이용하여 검색 할 수 있다. 질의 구문 분석은 검색 질의어로 들어온 구문을 분석하여 PDB플랫파일을 추출한다. 구조 질의 처리는 질의 구문분석을 통해 얻은 PDB플랫파일을 단백질 하위구조 생성모듈로 이동하여 SQL을 이용하여 단백질 하위구조 데이터베이스에 접근한다. 검색 결과 브라우징은 사용자의 질의를 받아서 하위구조비교와 주석데이터검색을 진행한 후 다시 웹을 통하여 결과를 보여준다.

3.2 단백질 하위구조 처리 모듈

단백질 하위구조 처리 모듈은 구조질의처리 컴포넌트와 공개용 단백질 구조 데이터베이스에서 추출한 구조 파일에서 필요한 구조정보만 파싱하여 하위구조를 정의하여 거리기반 매트릭스를 구축하여 새로운 플랫파일을 생성한다. 새로운 플랫파일에서 RMSD로 구조유사도를 비교한다.

3.2.1 구조 정보 파서

PDB파일로부터 활성 사이트를 포함하는 표면정보만을 얻기 위해 Rasmol을 이용하여 표면 아미노산 잔기들만을 얻는다. 우리 시스템은 단백질 구조로부터 얻을 수 있는 가장 기본이 되는 정보인 골격을 형성하는 Cα원자와 단백질 기능을 결정하는 아미노산 곁사슬 C_β의 좌표정보만을 사용하기 때문에 원본 PDB플랫파일에서 하위구조검색에 필요한 데이터들만 추출한다.

3.2.2 하위구조 생성

추출된 C_α와 C_β원자의 X, Y, Z 좌표 값과 잔기 서열번호를 이용하여 임의로 추출한 3개의 잔기들을 C_α끼리 연결하고 C_β 끼리 연결하여 하위구조를 생성한다. 이 때 세 개의 C_α원자 각 x,y,z 좌표를 하나의 점으로 보고 삼각형으로 표현될 수 있는 모든 경우의 수를 고려하여 추출하고 C_α삼각형 각 변의 길이에 사용자가 threshold값을 주어 표현할 수 있는 삼각형 수에 제한을 준다. 생성된 하위구조 정보를 이용하여 표 1처럼 새로운 플랫파일을 생성한다.

<표 1> 하위구조 파일 포맷 정의

PdbId	PDB 파일 식별자
TriId	하위구조 식별자
ChainID	체인 식별자
AresNum	하위구조 첫번째 잔기 번호
AresName	하위구조 첫번째 잔기 이름
BresNum	하위구조 두번째 잔기 번호
BresName	하위구조 두번째 잔기 이름
CresNum	하위구조 세번째 잔기 번호
CresName	하위구조 세번째 잔기 이름
Len1	첫번째 잔기의Ca원자와 첫번째 잔기의Ca원자의 거리
Len2	두번째 잔기의Ca원자와 세번째 잔기의Ca원자의 거리
Len3	세번째 잔기의Ca원자와 첫번째 잔기의Ca원자의 거리
Len4	첫번째 잔기의Ca원자와 두번째 잔기의Cβ원자의 거리
Len5	첫번째 잔기의Ca원자와 세번째 잔기의Cβ원자의 거리
Len6	두번째 잔기의Ca원자와 첫번째 잔기의Cβ원자의 거리
Len7	두번째 잔기의Ca원자와 세번째 잔기의Cβ원자의 거리
Len8	세번째 잔기의Ca원자와 첫번째 잔기의Cβ원자의 거리
Len9	세번째 잔기의Ca원자와 두번째 잔기의Cβ원자의 거리
Len10	첫번째 잔기의Cβ원자와 첫번째 잔기의Cβ원자의 거리
Len11	두번째 잔기의Cβ원자와 세번째 잔기의Cβ원자의 거리
Len12	세번째 잔기의Cβ원자와 첫번째 잔기의Cβ원자의 거리

3.2.3 거리기반 매트릭스

하나의 하위구조인 삼각형 두 개에서 표현된 6개의 점들을 이용해 모든 거리를 구한다. 하위구조에서 계산될 각각

의 C_α와 C_β 원자의 모든 거리 값을 매트릭스형태로 생성한다. 각 잔기들의 서로에 대해 유사성을 기술하는 테이블이라고 생각할 수 있다. 두 개의 삼각형으로 이루어진 점들의 거리 (Euclidean Distance)는 Cα원자 좌표 x, y, z와 그에 대응하는 3개의 C_β의 원자 좌표 x, y, z를 이용하여 계산한다. 이중 C_α원자와 이에 대응하는 C_β로 계산된 3개의 거리는 대부분 단백질에서 7.5Å 정도로 비슷하기 때문에 거리기반 매트릭스에 포함하지 않는다. 거리기반 매트릭스는 표 2처럼 12개의 거리를 구성한다.

<표 2> 거리기반 하위구조 파일 예제

193L	193L	194L	193L	1AT5	193L	3LYM	194L
5:04:07	15:14:13	14:13:16	15:16:13	16:13:18	14:15:16	19:18:16	20:18:16
A	A	A	A	A	A	A	A
5	15	14	15	16	14	19	20
ARG	HIS	ARG	HIS	GLY	ARG	ASN	TYR
4	14	13	16	13	15	18	18
GLY	ARG	LYS	GLY	LYS	HIS	ASP	ASP
7	13	16	13	18	16	16	16
GLU	LYS	GLY	LYS	ASP	GLY	GLY	GLY
3.79	3.80	3.80	3.81	5.16	3.80	3.80	5.47
4.98	3.80	5.16	5.16	5.25	3.81	5.71	5.71
5.50	5.28	5.47	5.28	5.71	5.47	7.94	6.76
2.42	2.44	2.45	4.71	4.50	4.61	2.45	4.48
5.67	4.31	6.52	4.31	6.05	6.52	6.78	5.43
4.42	4.61	4.39	2.44	5.36	2.44	4.53	6.58
5.54	2.45	5.36	4.50	6.56	4.71	4.75	4.75
4.51	5.34	4.53	5.34	4.75	4.53	7.57	8.24
4.48	4.39	4.50	5.36	6.09	2.44	6.05	6.05
2.99	3.30	3.01	3.31	5.12	3.30	3.16	5.57
4.84	3.01	5.12	5.12	7.21	3.31	5.02	5.02
4.66	4.49	5.69	4.49	5.02	5.69	6.34	6.88

3.2.4 하위구조 비교처리

데이터베이스에 저장된 하위구조들과 Root Mean Square Distance를 이용하여 구조유사도 비교한다. RMSD는 점들 사이의 기하학적 유사성을 결정하는 가장 기본적인 방법으로 점들 사이의 1대1 비교를 필요로 한다. A의 점들과 B의 점들 사이 거리의 평균을 더한 것의 평균값으로 정의된다.

3.3 단백질 하위구조 저장 관리 모듈

단백질 하위구조 저장 관리 모듈은 공개용 단백질 구조 데이터베이스에서 추출한 구조파일은 하위구조처리 모듈을 통해 생성된 하위구조플랫파일을 저장하고 그에 필요한 구조주석데이터는 주석정보데이터베이스에 입력된다. 이 모듈의 하위 컴포넌트들은 다음과 같다. 하위구조 데이터 저장기는 단백질 하위구조 생성 모듈 통해 생성된 하위구조들을 데이터베이스에 저장한다. 구조 주석 정보 저장기는 하위 구조비교의 결과와 함께 비교된 단백질의 주석정보도 저장하여 인터페이스에 보여줌으로써 단백질 구조 유사 부분뿐만 아니라 주석정보에 대해서도 비교 분석할 수 있도록 한다. 하위구조 데이터 자동 갱신기는 새로 배포된 PDB 엔트리들을 주기적으로 다운로드하여 구조데이터를 추출하고 새로운 플랫파일형식에 맞추어 하위구조 데이터베이스를 구축한다.

4. 결론

이 논문에서 우리는 웹 기반으로 단백질의 3차구조를 하위구조데이터베이스와 비교하여 새로운 구조가 이미 알려진 임의의 구조와 유사성을 가지고 있는지 검색한다. 따라서 기존 단백질의 구조정보도 주석데이터베이스로 시스템에 구축하여 관련구조 정보를 사용자에게 보여준다.

이 시스템은 PDB로부터 단백질의 플랫파일을 다운로드하고 하위구조비교프로그램에 필수적인 C_α원자와 C_β원자의 정보를 데이터를 이용하여 거리기반 매트릭스로 새로운 플랫파일형태로 저장한다. 그리고 기타 구조관련 필요한 데이터는 주석데이터베이스에 구축한다. 그리하여 구조비교와 단백질 연구에 필요한 데이터는 보다 빠르게 검색되고 사용자가 검색할 때 구조 정보결과가 상세하게 보여질 수 있다. 향후연구로는 단백질 하위구조에서 매치되는 부분을 시각적으로 보여줌으로서 보다 직관적으로 분석할 수 있도록 시스템을 업그레이드할 예정이다.

참고문헌

[1] Kwang Su Jung, Ki Jin Yu, Keun Ho Ryu, Yong Je Chung, "Predicting Ligand Binding Site using Protein Surface Features," PACIFIC SYMPOSIUM ON BIOCOMPUTING, pp.72, Jan. 2007.

[2] H.MBerman, J.Westbrook, Z.Feng, G.Gilliland, T.N.Bhat, H.weissing, I.N.Shindyalov, P.E.Bourne, "The Protein Data Bank", Nucleic Acids Research, Vol.28, pp.235-242, 2000

[3] Liisa Holm and Chris Sander "Protein structure comparison by alignment of distance matrices", Journal of Molecular Biology Vol.233,1993

[4] In-Geol Choi "Local feature frequency profile : A method to measure structural similarity in proteins" PNAS, Vol.101, pp.3797-3802, 2004

[5] Fabrizio Ferre, Gabriele Ausiello, Andreas Zanzoni and Manuela Helmer-Citterich, "SURFACE : a Database of Protein Surface Regions for Functional Annotation," Nucleic Acids Research, Vol. 32, pp.240-244, Jan. 2004.

[6] Nam Hee Yu, Kwang Su Jung, Yong Je Chung, Keun Ho Ryu, "Predicting Protein Function using Surface Comparison on Binding Area", Korean Society for Bioinformatics and Systems Biology, pp.153-154, Nov, 2008

[7] T. Andrew Binkowski, Larisa Adamian and Jie Liang, "Inferring Functional Relationships of Proteins from Local Sequece and Spatial Surface Patterns," J.Mol.Biol., Vol. 332, pp.505-526, Sep. 2003.