단백질 구조기반 단백질 간의 기능관계 예측 기법

김선신, 정광수, 류근호
충북대학교 전자계산학과
데이터베이스/바이오인포메틱스 연구실
e-mail:sskim04@chungbuk.ac.kr

A Method to Predict Protein Functional Relationships from Protein Structures

Sunshin Kim, Kwang Su Jung, Keun Ho Ryu
Dept of Computer Science, Chungbuk National University
Database/Bioinformatics Laboratory

요 약

단백질 구조로부터 단백질사이의 기능관계를 유추하는 일은 생명정보학에 있어서 중요한 연구과제이다. 여기서, 단백질 1차 구조로부터 단백질 기능관계의 예측이 용이한 진화적으로 가까운 종간에는, 아미노산 서열을 비교하여 결과를 획득하고, 진화적으로 먼 종간에는 단백질 3차 구조 및 표면구조를 종합적으로 활용함으로써, 단백질간의 기능관계를 보다 효율적이고 정확하게 예측할 수 있음을 보인다.

1. 서론

단백질 구조로부터 단백질 기능을 예측하고자 하는 연구는 생명정보학에 있어서 중요한 이슈가 되어왔다. 여기서 단백질 구조라면 다음 세 가지 경우로 나눌 수 있다. 단백질 서열, 3차원 단백질 구조와 단백질 표면으로 나타낼 수 있다. 단백질 서열로부터 단백질 기능을 예측하고자 하는 시도는, 풍부한 서열 데이터를 활용한 다양한 알고리즘들이 고안되어, 활발히 진행되어왔다. 그러나 진화적으로 거리가 먼 종간에는 서열 유사도가 매우 낮아서 그들로부터 상동성을 예측하는 일이 매우 어려운 것으로 여겨지고 있다. 따라서 이 문제를 해결하기 위한 시도로서 단백질 3차원 구조로부터 단백질 기능 관계를 유추하고자 하는 노력들이 있어왔다. 그러나 이 경우에도 완전한 예측이 불가능한 것으로 판명되어 단백질 표면으로부터 직접 기능관계를 유추하고자 하는 연구가 최근 활발히 진행되고 있다. 이 논문에서, 단백질 사이의 기능관계를 유추하고자 할 때, 단백질 구조의 특성을 종합적으로 고려함으로써 보다 빠르고 정확한 해결방법을 제시할 수 있음을 보이고자 한다.

이 논문은 2005년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음.

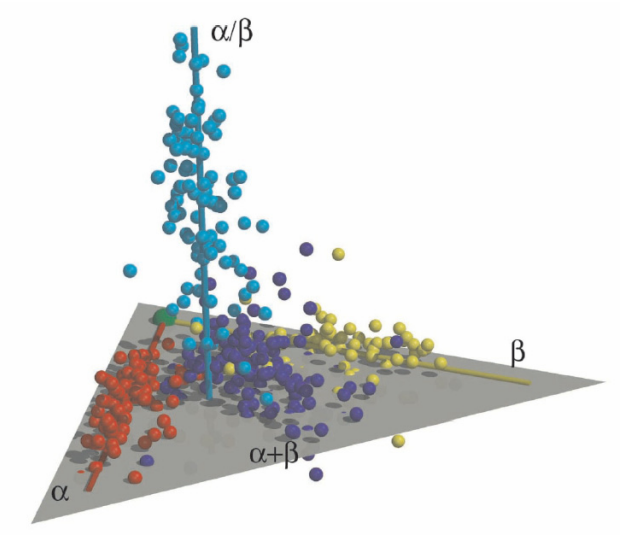
2. 관련연구

단백질 서열에 따라서 상동성을 결정하고 상동성에의해서 단백질의 기능을 유추하는 일은 생명정보학에 있어서 기본적이 패러다임이다. 하지만 단백질 서열 유사도가 40% 아래로 떨어지면 단백질사이의 기능예측이 매우 어려운 것으로 나타났다[1]. 이때 서열비교를 위해 사용하는 대표적인 도구가 BLAST[2]이다. 이는 휴리스틱(heuristic) 알고리즘을 사용하여 매우 효율적이고 빠른 반면에 서열유사도가 작은 경우에는 정확도가 많이 떨어지는 단점이 있다[3, 4, 5, 6].

이에 대한 해결책으로서, 서열 유사도가 높지 않은 경우라도 단백질 3차 구조의 비교로 단백질 기능관계를 식별할 수 있다는 사실이 밝혀졌다[7]. 3차 구조에서 구조와 구조간의 유사도를 비교하는 유명한 도구로는 VAST(Vector Alignment Search Tool)[8, 9]가 있는데 서열비교에 의해 탐지되지 않는 거리가 먼 상동성을 지닌 단백질을 빠른 시간에 찾을 수 있다는 장점이 있다.

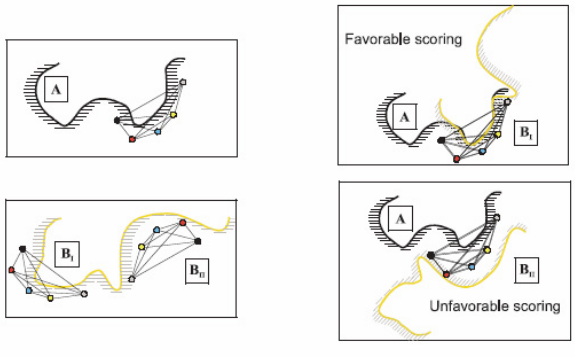
그러나 이 접근방식은 단백질 기능이 적은 수의 단백질 잔기 속에 암호화되어 있는 경우에는 정확한 결과를 얻는데 한계를 가지고 있다. 단백질 구조에서 유사한 폴드(fold)나 서열을 가지고 있다고 하더라도 완전히 서로 다른 기능을 가질 수 있음이 확인되었다[10]. 또한 단백질 상호간 분명한 기능적 관계를 가지고 있음에도 불구하고 구조 및 서열의 유사도와는 관련성이 없는 경우가 밝혀졌다[11].

이런 문제로 인하여, 단백질의 상호작용이 직접적으로 일어나는 단백질 표면으로부터 기능관계를 탐지하려는 연구들이 최근에 활발히 진행되고 있다. 한편 Hou등[12]은 (그림 1)에서 보여주는 것처럼 단백질 폴드공간에서 α, β, α/β, α+β 클래스로 나누어서 전범위에 걸친 단백질 구조 분포를 조사하였다.



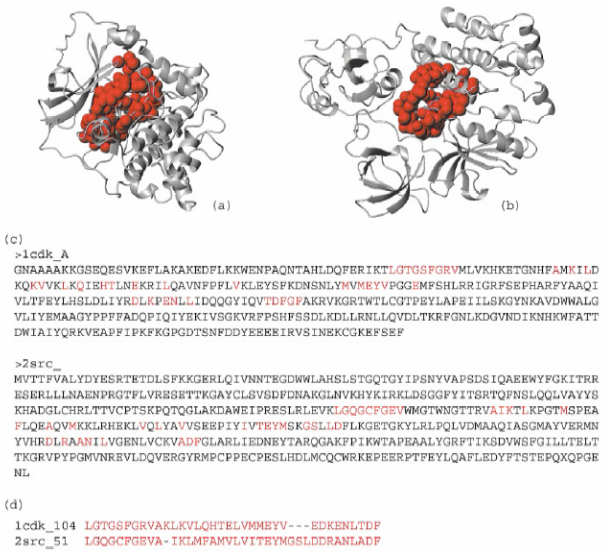
(그림 1) 단백질 폴드 공간의 3차원표현

Schmitt등[13]은 단백질 서열이나 폴드상동성과는 독립적으로 단백질 사이의 기능적 관계를 탐지하기 위한 새로운 방법을 개발하였다. (그림 2)에서 보여주는 것처럼 두 활성사이트(binding site)의 매칭정도를 측정하여 점수를 매기는 안을 고안하였다. 각각의 활성사이트에 있는 모든 원자를 고려하는 것이 아니고, 움푹한 구멍(cavity)을 정의하는 수도센터(pseudocenter)를 설정하여 단백질의 특정모양을 비교하였다. 활성사이트안에 있는 하위그룹 B_I와 B_{II}는 클릭 알고리즘(clique algorithm)에 의해 탐지되었다



(그림 2) 두 활성 파킷(binding pocket)의 매칭(matching)과 스코어링(scoring)

BinKowski등[14]은 단백질표면의 활성사이트에 해당하는 아미노산서열과 공간적 패턴을 탐지하여 단백질 기능관계를 유추하는 새로운 접근방법을 설명하였다. (그림 3)에서 두 단백질에서 활성사이트를 비교하면 서열 유사도(sequence identity)가 51%에 이르지만 단백질 전체서열 비교에서는 16%의 유사도를 갖는 것을 확인하였다. 또한 pvSOAR프로그램을 사용하여 단백질표면의 유사도를 구조적으로 측정하였다.

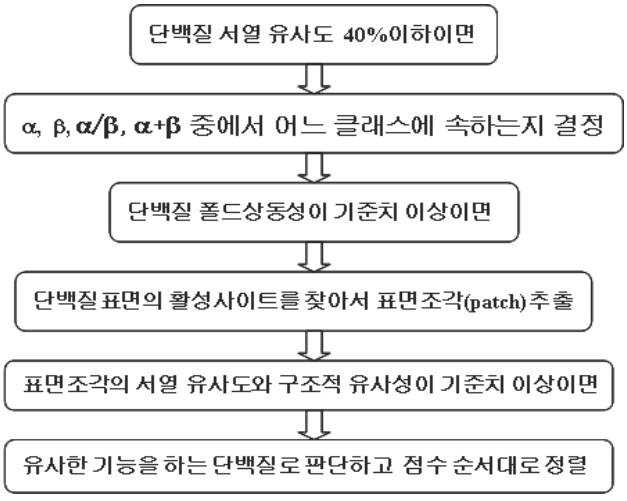


(그림 3) 두 단백질의 활성사이트의 서열비교

3. 제안한 방법

지금까지 살펴본 관련연구를 바탕으로 다음과 같은 방법을 제안한다. 이제 비교하고자 하는 새로운 단백질 A가 있고 데이터베이스에는 수많은 N개의

단백질들이 저장되어 있다고 하자. 이 경우 서로 다른 단백질사이의 기능관계를 알고자할 때, 크게 두 가지로 나누어서 고려하고자 한다. 첫째, BLAST프로그램을 사용하여 아미노산서열(단백질서열) 유사도가 40%이상이 되는지 아닌지를 테스트하여, 40%이상이면 BLAST프로그램의 결과를 그대로 사용하거나 또는 성능이 더 좋은 서열비교 프로그램을 사용함으로써, 기능관계 추측을 빠르고 정확하게 할 수 있다. 둘째, 만약 찾고자하는 단백질의 서열유사도가 40%이하가 되면, (그림 4)에서처럼 네 개의 단백질 폴드공간 중에서 어느 클래스에 속하는지를 Hou등[12]이 사용한 방법으로 알아낸다. 다음으로는 데이터베이스에 있는 모든 단백질을 다 비교하는 것이 아니라 찾고자하는 클래스에 해당하는 부분만을 검색함으로써 상당한 시간을 줄일 수 있다. 이때 VAST나 또는 그 외의 폴드상동성을 찾는 도구를 사용하여, 단백질구조를 비교해서 기준치 이상의 유사한 상동성을 가진 S개의 단백질을 추출한다. 이렇게 얻은 단백질 S개 중에서 단백질 A와 긴밀한 기능을 가진 단백질의 추출을 확실히 하기위해서, 획득한 S개의단백질 표면의 활성사이트를 찾아서 단백질 A의 활성사이트와 비교한다. 이때 단백질 A의 활성사이트 조각(patch)과 획득한 S개 단백질의 활성사이트 조각(patch)의 구조적 유사성 및 서열 유사성을 비교하여 기준치 이상의 값을 가진 단백질을 점수 높은 것부터 차례로 정렬한다.



(그림 4) 서열 유사도가 40%이하인 경우 유사한 기능을 하는 단백질을 탐색하여 획득하는 방법

4. 결 론

단백질 구조를 바탕으로 단백질사이에 기능관계를 유추하고자 하는 연구가 활발히 진행되고 있다. 1차원 단백질 구조에 해당하는 아미노산서열의 상동성에 의해서 단백질기능관계를 추측하려는 시도가 있었지만 진화적으로 거리가 먼 종들의 경우에 서열유사도가 작아서 정확한 예측을 하기가 매우 어려운 일로 판명되어졌다. 이로 인해서 3차원 단백질 구조의 폴드의 상동성에 근거하여 단백질기능관계를 유추하려는 노력이 있어왔다. 단백질 서열로는 기능예측을 명백히 밝힐 수 없는 경우에도 이 3차구조의 비교가 성공적인 결과를 줄 수 있음이 확인되었다. 그러나 이경에도 완전히 모든 경우를 다 충족시킬 수 없음이 입증되었고, 이제는 단백질 사이의 기능의 직접적 대상인 단백질표면의 구조적 특징에 기반 하여 단백질사이의 기능관계를 유추하고자 하는 노력들이 진행되고 있다. 이와 같은 연구결과들을 고려해보면, 진화적으로 거리가 가까운 종사이의 단백질 기능관계를 유추하고자 할 때는 단백질 1차구조만으로도 충분하므로, 단백질서열 유사도를 비교하는 도구만으로도 빠르고 손쉽게 결과를 얻을 수 있다. 그러나 서열유사도가 먼 종인경우는 1차구조만으로는 충분치 않으므로, 단백질 3차 구조와 단백질표면의 구조 및 서열을 종합적으로 고려함으로써 빠르고 정확한 결과를 얻을 수 있는 방법을 제안하였다.

참고문헌

[1] C. Wilson, J. Kreychman, & M. Gerstein ″Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores″ J. Mol. Biol. 297, 233-249, 2000

[2] S. F. Altschul, et al., "Basic local alignment search tool", J. Mol. Biol., Vol. 215, pp. 403-410, 1990

[3] A. R. Mushegian, et al., "Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes", Genome Res., Vol. 8, pp. 590-598, 1998

[4] M. Kanehisa & B. Peer, "Bioinformatics in the post-sequences era", nature genetics supplement, Vol. 33, pp. 305-310, 2003

[5] M. Y. Galperin & E.V. Koonin, "Sources of

systematic error in functional annotation of genomes: domain rearrangement, nonorthologous gene displacement and operon disruption", In Silico Biol., Vol. 1, pp. 55-67, 1998

[6] S. Kimmen, "Phylogenonmic inference of protein molecular function: advances and challenges", Bioinformatics, Vol. 20, No.2, pp. 170-179, 2004

[7] L. Holm and C. Sander, "Mapping the protein universe. Science, 273, 595-603, 1996

[8] J. F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison", Curr Opin Struct Biol, 6(3):377-385, 1996

[9] T. Madej, J. F. Gibrat, and S. H. Bryant, "Threading a database of protein cores. Proteins", 23(3):356-3690, 1995

[10] L. M. Kauvar and H. O. Villar, "Deciphering cryptic similarities in protein binding sites", Curr. Opin. Biotechanol., 9, 390-394, 1998

[11] A. Via, F. Ferre, B. Brannetti, A. Valencia, and M. Helmer-Citterich, "Three-dimensional view of the surface motif associated with the P-loop structure: *cis* and *trans* cases of convergent evolution", J. Mol. Biol., 303, 455-465, 2000

[12] J. Hou, G. E. Sims, C. Zhang, and S. Kim, "A global representation of the protein fold space", Proc. Natl. Acad. Sci. USA, 100, 2386-2390, 2003

[13] S. Schmitt, D. Kuhn, and G. Klebe, "A new method to detect related function among proteins independent of sequence and fold homology", J. Mol. Biol., 323, 387-406, 2002

[14] T. A. Binkowski, L. Adamian, and J. Liang, "Inferring functional relationships of proteins from local sequence and spatial surface patterns", J.Mol.Biol., 332, 505-526, 2003