MCL 알고리즘을 이용한 단백질 표면의 바인딩 영역 분석 기법

정 광 수[†] · 유 기 진[†] · 정 용 제[††] · 류 근 호[†††]

요 약

단백질은 다른 물질과의 결합하여 기능을 수행하기 때문에 활성 사이트가 유사한 단백질은 유사한 기능을 가진다. 따라서 단백질의 바인딩 영역을 식별함으로써 단백질의 기능을 추론할 수 있다. 이 논문은 MCL (Markov Cluster) 알고리즘을 이용하여 단백질의 바인딩 영역을 추출하는 새로운 방법을 제시한다. 이를 위하여 단백질의 표면 잔기 거리를 나타내는 distance matrix를 생성하고, 여기에 MCL 프로세스를 적용한다. 제시한 방법을 평가하기 위해 Catalytic Site Atlas (CSA) 데이터를 사용하였다. CSA 데이터 (94개의 단일 체인 단백질)를 이용한 실험 결과, 알고리즘은 91개 단백질의 활성 사이트 주변의 바인딩 영역을 검출하였다. 이 논문은 단백질 활성 사이트를 분석하기 위한 새로운 기하학적 특징을 제시하였고, 활성 사이트와 관련이 없는 잔기를 제거함으로써 단백질 표면의 분석의 시간을 줄일 수 있는 장점이 있다

키워드 : 단백질 구조, 표면, 바이오인포매틱스, MCL 알고리즘

Investigating Binding Area of Protein Surface using MCL Algorithm

Kwang Su Jung[†] · Ki Jin Yu[†] · Yong Je Chung[††] · Keun Ho Ryu[†††]

ABSTRACT

Proteins combine with other materials to achieve their function and have similar function if their active sites are similar. Thus we can infer the function of protein by identifying the binding area of proteins. This paper suggests the novel method to select binding area of protein using MCL (Markov Cluster) algorithm. We construct the distance matrix from surface residues distance on protein. Then this distance matrix is transformed to connectivity matrix for applying MCL process. We adopted Catalytic Site Atlas (CSA) data to evaluate the proposed method. In the experimental result using CSA data (94 selected single chain proteins), our algorithm detects the 91 (97%) binding area near by active site of each protein. We introduced a new geometrical features and this mainly contributes to reduce the time to analyze the protein by selecting the residues near by active site.

Key Words : Protein Strucure, Surface, Bioinformatics, MCL Algoritm

1. 서 론

단백질은 생명체에서 주요 핵심 역할을 담당한다. 생명체의 생물학적 기능과 과정을 밝히기 위해서 단백질은 다양한 방법으로 분석되었다. 단백질 분석 방법 중 단백질 서열 분석을 이용한 상동성 비교는 쉽고 빠르게 비교할 수 있지만, 상동성이 낮은 경우는 예측이 어렵다. 서열 분석을 이용하여 예측이 불가능한 경우는 폴드 분석을 실행하게 된다. 그러나 기능이 유사하더라도 폴드가 다른 단백질이 있고[1], 유사한 폴드나 서열을 갖는 단백질도 완전히 다른 기능을 수행하기도 한다[2]. 따라서 이런 문제점을 보완하기 위해 기능과 밀접한 관계를 가지고 있는 특정 핵심 영역의 연구가 필요하다[3]. 단백질 기능은 대부분 단백질 표면의 물리적, 화학적, 기하학적 특징 등에 의해 결정되며, 유사한 폴드라고 할지라도 단백질 표면의 활성 사이트가 다른 경우 다른 기능을 수행할 수 있다. 단백질 표면의 서열과 공간적 패턴 분석을 통하여, 핵심 잔기와 단백질 기능과의 관계를 규명함으로써 알려지지 않은 단백질 구조의 기능을 밝힐 수 있고, 단백질 표면과 기능의 새로운 관계를 발견할 수 있다. 단백질은 기질, 리간드, RNA, 단백질 등의 물질과 결합하여 기능을 수행하고, 이 중 효소는 리간드와 결합하는 단백질의 한 종류이다. 효소의 몇몇 아미노산은 효소의 촉매 작용이 일어나는 사이트를 구성하기 때문에, 활성 사이트는 특정 기질의 리간드와 결합하는 아미노산을 갖는다. 이 논문은 효소를 대상으로 연구를 진행하였으며, 바인딩 영역은 리간드와 결합하는 활성 사이트 주변의 잔기들을 의미한다.

※ 이 논문은 2007년 교육인적자원부의 재원으로 한국학술진흥재단의 지원(지방연구중심대학육성사업/충북BIT연구중심대학육성사업단)과 과학기술부의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(R01-2007-000-10926-0).
† 정 회 원 : 충북대학교 전자계산학과 연구원
†† 정 회 원 : 충북대학교 생명과학부 교수
††† 종신회원 : 충북대학교 전기전자컴퓨터공학부 교수(교신저자)
논문접수 : 2007년 7월 27일, 심사완료 : 2007년 10월 27일

우리는 인접한 잔기 간의 연결 형태를 이용하여 단백질 표면에 있는 활성 사이트 주변의 바인딩 영역을 검출하는 새로운 방법을 제시한다. 그리고 잔기의 거리 정보를 나타내는 distance matrix를 변환하여 connectivity matrix를 생성한다. connectivity matrix에서, 사용자가 정의한 거리 cut-off threshold를 만족하지 못하는 값은 무시되고 '0'으로 대체된다. 이 matrix는 MCL(markov cluster) 알고리즘의 입력 matrix로 이용된다[4,5]. 우리는 weight factor를 설정하기 위해 Catalytic Site Atlas (CSA, http://www.ebi.ac.uk/thornton-srv/databases/CSA/) [6]를 사용하였다. CSA 데이터를 이용하여 활성 사이트 주변의 바인딩 영역에서 빈번하게 발견되는 잔기 빈발도를 계산하고, 바인딩 영역에서 빈발하게 발견되는 잔기는 높은 가중치를 할당한다. 잔기들의 가중치에 대한 설명은 4장에서 자세히 설명할 것이다. 우리는 distance matrix에서 의미 있는 거리 값을 선택하기 위한 distance cut-off 값과 MCL 알고리즘의 다양한 inflation factor를 가지고 테스트하였다. CSA 데이터(94개 단일 체인 단백질)를 이용하여 실험한 결과, 우리의 방법은 91개 단백질에서 활성 사이트 주변의 바인딩 영역을 검출하는 결과를 보였다. 각 단백질에서 선택된 가장 좋은 클러스터들을 보면, 대략 70% 잔기가 실제로 바인딩 영역에 속하는 것으로 나타났다. 다른 실험 결과와 분석은 5장에서 자세히 설명될 것이다. 우리의 연구는 단백질 표면의 활성사이트에 관련된 새로운 기하학적 특징을 제시하였고, 이를 이용하여 활성 사이트와 관련 없는 잔기를 필터링하고, 활성 사이트 주변의 잔기를 선택함으로써 활성 사이트 분석 시간을 줄일 수 있다. 또한, 기존의 바인딩 영역이 오목한 형태가 아니더라도 제안한 방법을 적용하면, 바인딩 영역예측이 가능하다.

2. 관련연구

단백질 서열 분석과 구조 분석은 서열과 구조의 유사성이 낮음에도 불구하고 유사한 기능을 수행하는 단백질의 경우 기능을 정확하게 예측할 수 없다[7, 8]. 따라서 구조와 특징에 의해 단백질 기능을 결정하는 활성 사이트의 분석이 많이 이루어지고 있다. 이 절에서는 활성 사이트를 분석하여 단백질의 기능을 예측하는 기법들을 소개한다.

활성 사이트의 구조를 분석한 Fabian Glaser[9]는 SURFNET[10] 프로그램을 통해 표면의 가장 큰 포켓(pocket) 4개를 선택하고, 그 중 활성 사이트를 포함하는 포켓을 식별하는 방법을 제안하였다. 활성 사이트를 포함하는 포켓은 결합 물질보다 크며, 효소는 기질과 보조인자 등 하나 이상의 물질과 결합한다. 따라서 정확한 활성 사이트 크기를 이용하여 결합물질을 밝히기 위해 Consurf-HSSP(Homology-Derived Secondary Structure of Proteins) 데이터베이스에서 잔기의 보존율을 측정하고, 보존율이 높은 중요한 잔기와 거리가 먼 잔기를 제거하였다. 이 방법은 SURFNET의 초기 포켓 부피와 아미노산 잔기 수를 감소시켜 더 정확한 활성 사이트의 형태와 포켓 내의 위치를 예측한다.

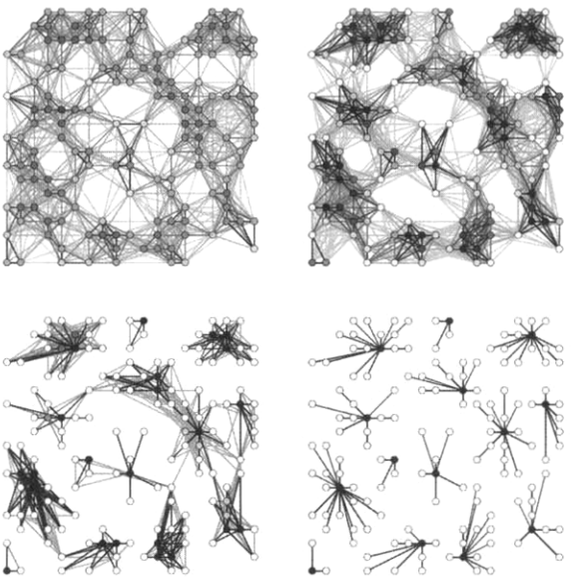
Changhui Yan[11]은 구조 정보 없이 활성 사이트 아미노산 잔기와 서열상의 주변 아미노산 잔기를 이용하여 분석하였다. 또한 DSSP(Database of Secondary Structure in Proteins) 프로그램을 이용하여 단백질 표면의 면적을 나타내는 ASA(solvent Accessible Surface Area)를 측정하였다. 단백질 표면과는 다른 특징을 가지는 활성 사이트 아미노산 잔기의 성질과 측정한 ASA를 이용하여, 표면보다 빈발한 아미노산 잔기의 물리화학적 특징을 추출하였다. 추출된 특징은 지식기반 분류기(SVM : Support Vector Machine)를 통해 활성 사이트의 아미노산 잔기를 분류하는데 이용된다. 이 논문에서는 서열상의 주변 아미노산 잔기를 선택하지 않고 폴드상의 주변 아미노산 잔기를 이용한다.

T.Andrew Binkowski[12]는 CASTp(Computed Atlas of Surface Topography of proteins) 데이터베이스에서 포켓의 정보를 추출하여 분석하였다. 표면과 포켓을 구성하는 아미노산 잔기 특성을 분석한 결과, 포켓에서는 방향성 아미노산 잔기와 소수성 아미노산 잔기가 비교적 많이 분포하고, 표면에서는 극성 아미노산 잔기가 많이 분포한다. 그리고 포켓을 구성하는 아미노산 잔기의 서열 보존율은 다른 부분보다 더 높게 측정되었다. 포켓의 형태를 비교하기 위해, 아미노산 잔기의 위치를 나타내는 cRMSD와 잔기의 벡터를 나타내는 oRMSD를 계산하였다.

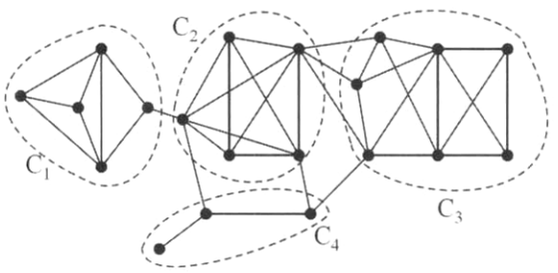
Oliviero Carugo[13]는 고유벡터(eigenvector)와 고유치(eigenvalue)를 이용하여 단백질 표면 패치(patch)를 분석하고, 중심 원자와 그 주변의 40개 원자로 구성된 패치의 고유벡터를 통해 단백질 표면을 비교하였다. 41개 원자의 x, y, z 좌표는 41×3 행렬 D로 표현하고, D의 전치행렬을 이용하여 3×3 행렬 Z로 변환한다. 행렬 Z를 통해 계산된 고유치는 41개 원자가 분산되어있는 정도를 나타내고, 각 패치의 고유치를 이용하여 패치의 형태를 비교하였다. Oliviero Carugo는 41개의 원자를 선택한 반면, 우리는 각 바인딩 잔기 주변의 아미노산 잔기 20개를 선택한다.

Susan Jones[14]는 일반적으로 단백질-단백질 상호작용의 결합 부위인 interface의 형태와 △ASA, 아미노산 잔기 특징, 소수결합 등의 정보를 분석하였다. △ASA는 monomer의 ASA와 complex 형성 후의 ASA 차이를 나타내는 것으로, homodimer와 heterocomplex의 △ASA 차이를 분석하였다. interface의 평평한 정도를 분석하기 위해 dRMSD를 계산한 결과, heterocomplex interface는 움푹 들어간 형태인 반면, homodimer는 울퉁불퉁한 형태임을 밝혔다. 표면에 분포하는 특정 아미노산 잔기의 ASA와 interface에 분포하는 특정 아미노산 잔기 ASA 비율은 특정 잔기 분포정도를 나타낸다. 이를 통해 interface에는 표면보다 소수성 아미노산 잔기가 많이 분포하는 것을 밝혔다. 이 논문에서는 Susan Jones의 물리화학적 특징을 계산하는 식을 변형하여 아미노산 잔기 분포 정도를 밝힌다.

이 논문에서 사용되는 MCL 알고리즘[4, 5]은 그래프에서 랜덤 워크와 flow 시뮬레이션을 기반으로 한 빠른 그래프 클러스터링 알고리즘이다. 특히, 간단한 그래프와 가중치가 있는 간선이 표기된 그래프를 다루기 위해 설계되어, 그래프 클러스터링 분야에서 사용되어왔다.



(그림 1) MCL process에 의한 flow 시뮬레이션 단계



(그림 2) connectivity graph G의 예

MCL process를 나타낸 (그림 1)에서 반복적인 단계는 왼쪽에서 오른쪽 방향으로, 위에서 아랫방향으로 이루어진다. 왼쪽 상단의 그래프는 MCL process 전의 처음 단계이고, 오른쪽 하단의 그래프는 flow가 한계에 이르러 MCL process가 종료된 상태이다. 검은 노드는 어트랙터(attractor)를 나타내며, 이는 검은 노드를 중심으로 다른 흰색 노드들을 끌어당기고 있는 것을 연결 형태를 나타내고 있다. 즉, 오른쪽 하단의 그래프는 MCL 알고리즘을 적용한 그래프 클러스터링의 결과를 나타내며, 연결된 노드들을 클러스터링한 것이다.

MCL 알고리즘은 expansion과 inflation 연산이 반복적으로 수행함으로써 랜덤 워크가 이루어진다. (그림 2)에서 각각의 노드들은 다른 노드와의 연결 정보를 가지고 있고, 가중치가 높은 노드는 높은 값을 가지고 있다. 이는 matrix 형태로 표현이 가능하다. 그림 2에서의 C_1, C_2, C_3 와 C_4는 MCL 알고리즘을 이용하여 클러스터링된 결과를 나타낸다. 그림 3은 MCL process의 단계이다.

Expansion은 표준 matrix 제곱(product)을 이용하여 확률 matrix의 제곱을 취하여 계산된다. (그림 3)에서의 확률 matrix (Markov Matrix, M)의 한 열(column)은 음수가 아닌 각 cell value로 구성되고, 한 열의 각 셀 값의 합이 1이다. $M \geq 0$을 만족하는 $M \in R^{kxk}$ matrix와 r >1의 real number를

1. 그래프 G를 이용하여 Graph Matrix GM 생성
2. GM을 확률적 Markov matrix M으로 변환
3. $M = M^2$ (Expansion)
4. $M = \Gamma_r M$ (Inflation)
5. 만약 $M \neq M^2$ 이라면, 3단계와 4단계 반복.
   $M = M^2$ 이라면, 중단.

(그림 3) MCL Process

이용하여, 제곱 계수 r을 열의 각 cell value에 취하여 산출된 한 열의 확률 matrix는 $\Gamma_r M$이다. Γ_r은 식 (1)에서 제곱 계수 r을 이용하는 inflation 연산자이다. Expansion 과 inflation 단계를 되풀이 하면서, 한 단계의 expansion 과 inflation 이 끝날 때 마다, matrix는 한 열의 셀의 합이 1인 확률 matrix 형태로 변형되고, 이를 다음단계의 MCL process에 이용한다.

$$(\Gamma_r M)_{pq} = (M_{pq})^r / \sum_{i=1}^{k} (M_{iq})^r \qquad (1)$$

확률 matrix M의 각 열 j는 M과 관련된 확률 그래프의 j 노드와 일치한다. j 열의 i 행 entry(즉, M_{ij})는 j 노드에서 i 노드로 흐르는 확률이다. r>1의 값은, 흐름 시뮬레이션의 시작을 나타내고, 쉽게 흐를 수 있는 노드와 덜 흐를 수 있는 노드를 구분하여 시뮬레이션 됨으로써 클러스터링이 진행된다.

3. 표면 잔기의 Distance Matrix

선행 연구[15,16,17,18,19]의 단백질 잔기 distance matrix는 단백질의 전체 잔기를 포함하고 단백질 폴드를 비교하기 위해 사용되었다. 이 논문에서 우리는 표면 잔기의 연결 정보를 나타내기 위해 단백질 표면의 잔기 간의 거리를 계산하였다. 기존의 단백질 distance matrix는 잔기의 Cα의 위치를 기준으로 distance matrix를 구성하는데 반하여, 우리는 Residue Center를 사용하였다. Ca의 위치는 결국 단백질 전체 구조의 백본을 나타내며, 기존의 distance matrix 기반의 선행 연구는 단백질 전체 구조를 사용하기 때문에 Ca의 위치를 사용하는 것이 타당하나, 우리의 경우는 단백질의 전체 구조 보다는 기능과 밀접한 연관이 있는 표면의 특정 영역을 대상으로 한다. Residue Center는 Cα 좌표를 포함한 side chain atom의 평균으로 정의된다. 대부분의 side chain이 단백질 내부보다 표면의 방향으로 위치하기 때문에 side chain을 포함한 Residue Center는 Cα 위치보다 더 표면과 근접한 형태를 나타낸다.

$$x \ coordinate \ of \ Residue \ Center = \frac{\sum_{i=1}^{n} (C\alpha_x + ax_1 + ax_2 + ax_3 + \cdots + ax_n)}{n+1} \qquad (2)$$

식 (2)은 Residue Center의 x 좌표 계산 방법이며, y와 z 좌표도 동일한 방법으로 계산된다. 식 (2)에서, Ca_x는 Cα의 x 좌표를 나타내고, ax_n은 각 side chain atom의 x 좌표를 의미한다. 우리는 표면 잔기 간의 Euclidean Distance를 계산하여 distance matrix를 생성하였다. 예로, (그림 4)의 각 cell value는 Residue Center 간의 거리를 나타낸다.

자기와의 거리를 나타내는 대각선 셀의 값은 '0'을 갖고, 1차 아미노산 서열에서 근접한 잔기는 공간상에서도 근접하였기 때문에 낮은 값 갖는다. (그림 4)의 distance matrix는 connectivity matrix로 변환된다. 여기서, 공간상에 서로 근접한 잔기는 높은 connectivity를 갖는다고 가정한다. 그리고 다양한 distance cut-off value를 적용하였으며, 적용 결과는 5절에서 소개한다. connectivity matrix는 MCL process의 input matrix로 사용된다. 다음 장에서 connectivity matrix의 변환에 대해서 더 자세히 설명한다.

	aa1	aa2	aa3	aa4	aa5	aa6
aa1	0	5	8	12	13	3
aa2	5	0	6	9	8	20
aa4	8	6	0	3	10	16
aa4	12	9	3	0	2	8
aa5	13	8	10	2	0	5
aa6	3	20	16	8	5	0

(그림 4) Distance Matrix of surface residues.

4. MCL 알고리즘의 응용

Distance matrix의 모든 셀 값 D_{ij}는 Graph Matrix GM을 생성하기 위해 $1/D_{ij}$로 변환된다. distance matrix에서는 낮은 cell value가 더 의미 있는 반면, Connectivity graph G의 Graph Matrix GM에서 높은 cell value가 더 중요한 의미를 갖는다. (그림 5)는 Graph Matrix GM로 변환된 distance matrix를 나타낸다.

	aa1	aa2	aa3	aa4	aa5	aa6
aa1	0	1/5	1/8	1/12	1/13	1/3
aa2	1/5	0	1/6	1/9	1/8	1/20
aa4	1/8	1/6	0	1/3	1/10	1/16
aa4	1/12	1/9	1/3	0	1/2	1/8
aa5	1/13	1/8	1/10	1/2	0	1/5
aa6	1/3	1/20	1/16	1/8	1/5	0

(그림 5) connectivity Graph Matrix GM로 변환된 distance matrix

그러나 여기서 (그림 5)의 대각선상의 '0' cell value를 어떻게 설정할 지의 문제가 남아있다. (그림 1)에서 언급한 것처럼, graph G의 attractor는 attractor 주변의 cluster를 생성한다. 여기서 우리는 바인딩 영역에서 빈번하게 발견되는 잔기들을 attractor로 정의하고 이들에게 높은 값의 가중치를 부여한다. 만약 대각선 셀에 더 큰 값을 할당한다면 이는 attractor 역할을 하게 되고, 다른 node(잔기)와 높은 connectivity를 가지게 되어 이를 중심으로 클러스터가 형성될 가능성이 크다. 우리는 대각선상의 셀 값에 일정한 값을 설정하기보다, 바인딩 영역의 잔기 빈발도를 계산하여, 대각선상의 값을 설정하였다.

4.1 바인딩 영역의 아미노산 빈발도

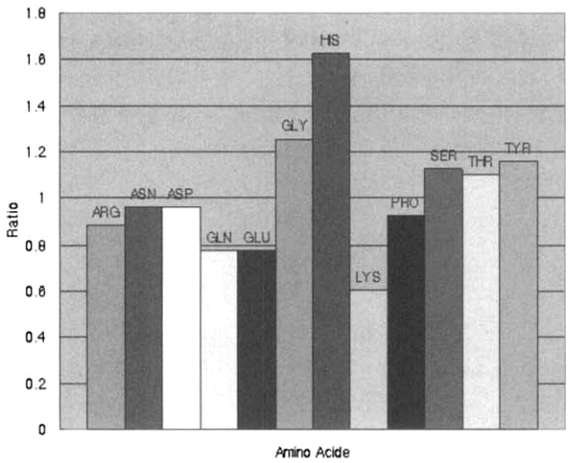
단백질 상호작용은 결합하는 잔기를 포함하는 더 넓은 영역에서 수행된다. 이 논문에서 결합하는 잔기 주변의 영역을 바인딩 영역으로 정의한다. 단백질의 활성 사이트를 포함하는 바인딩 영역은 단백질 표면에 위치하고, 잔기 구성은 단백질 표면 전체와는 다르다. 바인딩 영역의 잔기 빈발도를 계산하기 위해, 우리는 CSA [6] 데이터베이스로부터 94개의 단일 체인 단백질의 활성사이트 잔기를 추출하였다. 표면 잔기를 추출하기 위해 RasMol [20] 프로그램을 사용하였고, 각 활성사이트 잔기로부터 가장 근접한 20개의 잔기들을 선택하였다. 결국 선택된 잔기는 중복을 제거하고 바인딩 영역을 형성한다.

잔기 빈발도를 계산하는 과정에서 우리는 표면보다 바인딩 영역에서 더 빈발하게 발견되는 잔기를 조사하였다. 특히 GLY(Glycine)은 바인딩 영역에서 가장 빈발하게 발견된다. 그러나 GLY은 단백질 표면 전체에서도 가장 빈발하기 때문에 GLY의 빈발도 조정이 필요하다. 식 (3)은 아미노산 빈발도 AA_i의 계산 식이다.

$$AA_i = \frac{\left(\sum_{a=1}^{N} AA_i(a) / \sum_{a=1}^{N} AA(a)\right)}{\left(\sum_{s=1}^{N} AA_i(s) / \sum_{s=1}^{N} AA(s)\right)} \quad (3)$$

표면에서의 특정 잔기 빈발도 AA_i는 식 (3)의 분모에서 정의된다. AA(s)는 한 단백질에서 표면 잔기의 전체 개수를 나타내고, $AA_i(s)$는 특정 잔기 AA_i의 개수이다. 바인딩 영역에서의 특정 잔기 AA_i 빈발도는 식 (3)의 분자에서 정의된다. AA(a)와 $AA_i(a)$는 각각 바인딩 영역 잔기의 전체 개수와 바인딩 영역의 특정 잔기 AA_i 개수를 나타낸다.

(그림 6)는 식 (3)을 이용한 실험 결과를 나타낸다. (그림 6)에서 'AA_i = 1.0'은 아미노산 AA_i가 바인딩 영역과 표면에서 동일하게 빈발함을 의미한다. 'AA_i < 1.0'은 특정 잔기가 바인딩 영역보다 표면에서 더 빈발한 것을 의미한다. 반대로, 특정 잔기 AA_i가 표면보다 바인딩 영역에서 더 빈발한 경우 'AA_i > 1.0'의 결과를 가진다. 이 비율을 기반으로, 우리는 각 아미노산에 대한 가중치를 계산하고 (그림 5)의 대각선상의 값 설정 시 계산된 가중치를 사용한다.



(그림 6) 아미노산 빈발도

	aa1	aa2	aa3	aa4	aa5	aa6
aa1	W_{aa1}	1/5	1/8	1/12	1/13	1/3
aa2	1/5	W_{aa2}	1/6	1/9	1/8	1/20
aa4	1/8	1/6	W_{aa3}	1/3	1/10	1/16
aa4	1/12	1/9	1/3	W_{aa4}	1/2	1/8
aa5	1/13	1/8	1/10	1/2	W_{aa5}	1/5
aa6	1/3	1/20	1/16	1/8	1/5	W_{aa6}

(그림 7) diagonal value 이용한 connectivity Graph Matrix GM

실험 결과에서 알 수 있듯이, HIS (Histidine), GLY (Glycine), TYR (Tyrosine), SER (Serine)은 LYS (Lysine), GLU (Glutamic acid), GLN (Glutamine)보다 더 높은 가중치를 가진다. (그림 7)은 connectivity Graph Matrix GM으로 변환된 distance matrix를 보여주고, W_{aan}은 특정 아미노산에 대한 가중치를 나타낸다. (그림 8)은 이 논문에서 사용된 MCL process의 전체적인 단계이다.

5. Validation

우리는 94개의 단백질을 이용하여 MCL 알고리즘의 inflation factor와 3장에서 언급한 distance cut-off를 다양하게 설정하였다. 그리고 단백질 표면의 활성 사이트 주변에 위치하는 유효 클러스터를 생성하는 단백질의 수와 바인딩 영역에 위치하는 유효 클러스터의 precision을 계산하였다. 제안하는 방법의 검증을 위하여 다음과 같은 바인딩 영역의 정의를 요한다.

【정의1】 바인딩 영역

바인딩 영역은 활성사이트 잔기주변의 잔기를 지칭하는 용어로써 활성사이트 각 잔기 당 가장 가까운 20개의 잔기를 취한다. 단백질 마다 활성사이트를 이루는 잔기 수는 다양하므로, 바인딩 영역의 크기(잔기의 개수)를 고려할 때, 각 잔기별로 중복된 잔기들은 제거되어 카운트 된다.

5.1 DataSet

우리는 CSA [6] 데이터베이스로부터 94개 단일 체인 효소를 추출하였다. Catalytic Site Atlas (CSA)는 효소의 3차원 구조상의 촉매 잔기와 활성 사이트 정보를 제공하는 데이터베이스이다. CSA 데이터베이스는 효소가 일으키는 촉매 작용의 종류에 따라, 작용에 관여하는 촉매 잔기를 분류하였다. RasMol [20] 프로그램은 표면 잔기를 추출하기 위해 사용되었고, 각 바인딩 잔기와 가장 근접한 20개 잔기를 선택하였다. 중복을 제거한 잔기는 우리가 정의한 바인딩 영역을 형성한다.

5.2 유효 클러스터

유효 클러스터를 평가하기 위해 클러스터의 precision과 cluster score를 제시하였다. cluster score는 각 아미노산의 가중치로부터 계산된다. 바인딩 영역에서 빈발하는 잔기로 구성된 클러스터는 평균적으로 높은 가중치들을 갖는다(식 4).

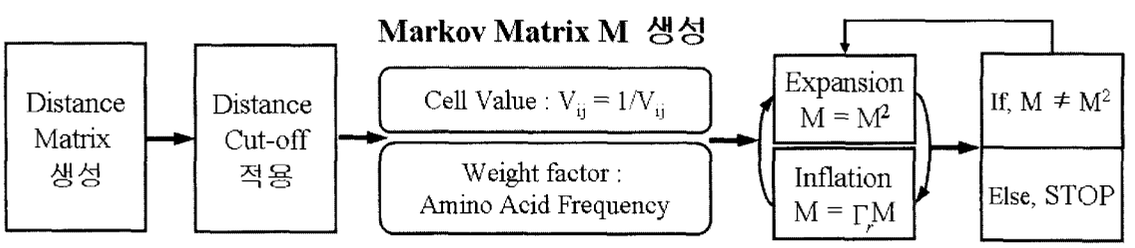
$$cluster\ score = \frac{\sum_{i=1}^{n}(W_{aa1} + W_{aa2} + \cdots + W_{aan})}{n} \quad (4)$$

【정의2】 유효 클러스터

MCL의 결과로서 제시된 클러스터 중 cluster score 가 높고, 클러스터의 precision이 높은 클러스터를 나타낸다. 각 클러스터의 cluster score는 식 (4)로 계산되고, precision은 식 (5)로 계산된다.

【정의3】 대표 클러스터

하나의 단백질은 inflation과 distance cut-off에 따라 여러 개의 유효 클러스터를 가질 수 있고 이중 가장 좋은 precision을 갖는 클러스터를 그 단백질의 대표 클러스터라



(그림 8) MCL 알고리즘을 적용한 프레임 워크

고 정의한다.

다른 평가 기준은 cluster의 precision이다. cluster의 잔기 중 바인딩 영역의 잔기는 true positive 잔기로 분류되고, 그 외의 잔기는 cluster에 의해 잘못 바인딩 영역의 잔기로 분류된 false positive이다. 따라서 바인딩 영역의 잔기를 많이 포함하는 cluster일 수록 높은 precision을 갖는다(식 5).유효 클러스터는 정의 2에 따라 best 또는 second cluster score와 특정 inflation factor 와 distance cut-off를 적용 했을 때, 가장 좋은 precision을 갖는 cluster를 의미한다. 94개 단백질 중, 우리가 제시한 방법은 best cluster score의 유효 클러스터를 갖는 68개 단백질(72%)을 산출하였다. <표 1>에서 best cluster score를 이용하여 검출된 68개 단백질의 'cluster score' 열은 'Best'로 나타내었다. Second cluster score를 갖는 cluster가 유효 클러스터로 분류된다면, 91개 단백질(97%)이 예측된다. Second cluster score을 이용하여 검출된 23개 단백질은 <표 1>의 'cluster Score'에서 값 'Second'로 나타내었다.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

우리는 9개의 distance cut-off value (11~19 Å)와 8개 inflation factor (1.2~1.9)를 조합하여 94개 단백질을 실험하

〈표 1〉 단일 체인 효소 DataSet의 94개 단백질 분류와 실험 결과

Ec num	Pdb code	Representative Precision	Active Site	Cluster Score	Description
Oxidoreductases : 15					
1.1.1.2	2alr	0.92	In	Best	Aldehyde reductase
1.1.1.158	1mbb	0.5	In	Second	Uridine Diphospho-N-Acetylenolpyruvylglucosamine
1.5.1.3	1ra2	0.93	In	Best	Dihydrofolate reductase
1.5.1.3	1vie	1.0	In	Best	Dihydrofolate reductase
1.5.1.28	1bg6	1.0	In	Best	N-(1-d-carboxylethyl)-l-norvaline dehydrogenase
1.7.2.1	1nid	1.0	In	Best	Nitrite reductase
1.7.3.3	1uox	0.7	In	Second	Urate oxidase
1.8.1.2	1aop	0.72	In	Best	Sulfite reductase hemoprotein
1.11.1.10	1a8q	0.9	In	Best	Bromoperoxidase A1
1.11.1.10	1vnc	1.0	In	Best	Vanadium-containing chloroperoxidase
1.11.1.10	2cpo	Not detected			Chloroperoxidase
1.13.12.7	1lci	0.53	In	Best	Luciferase
1.14.15.1	1akd	0.57	In	Second	Cytochrome p450cam
1.15.1.1	1eso	1.0	Out	Second	Cu, Zn superoxide dismutase
1.15.1.1	2jcw	0.81	In	Best	Cu/zn superoxide dismutase
Transferases : 17					
2.1.1.6	1vid	0.89	In	Best	Catechol o-methyltransferase
2.1.1.45	1lcb	0.91	In	Best	Thymidylate synthase
2.1.1.80	1af7	0.92	In	Best	Chemotaxis receptor methyltransferase cher
2.1.4.1	1jdw	0.68	In	Best	L-arginine-glycine amidinotransferase
2.3.1.39	1mla	1.0	In	Best	Malonyl-coenzyme a acyl carrier protein transacylase
2.3.1.41	1kas	0.42	Out	Second	Beta-ketoacyl acp synthase ii
2.3.1.117	2tdt	0.92	In	Best	Tetrahydrodipicolinate n-succinyltransferase
2.3.1.129	1lxa	0.92	In	Best	Udp n-acetylglucosamine o-acyltransferase
2.3.3.1	1al6	0.75	In	Best	Citrate synthase
2.4.2.29	1pud	0.83	In	Best	Tma-guanine transglycosylase
2.4.2.30	1a26	0.92	In	Best	Poly (adp-ribose) polymerase
2.5.1.7	1uae	1.0	Out	Second	Udp-n-acetylglucosamine enolpyruvyl transferase
2.5.1.15	1aj0	0.8	In	Best	Dihydropteroate synthase
2.7.3.3	1bg0	0.71	In	Second	Arginine kinase
2.7.4.3	1zio	1.0	In	Best	Adenylate kinase
2.7.6.3	1hka	0.94	In	Second	6-hydroxymethyl-7,8-dihydropterin
2.8.1.1	1rhs	0.94	In	Best	Sulfur-substituted rhodanese

〈표 1〉 계속

Ec num	Pdb code	Representative Precision	Active Site	Cluster Score	Description
Hydrolases : 42					
3.1.1.-	1agy	1.0	In	Best	Cutinase
3.1.1.3	2lip	0.68	Out	Best	Lipase
3.1.1.6	1bs9	1.0	In	Best	Acetyl xylan esterase
3.1.1.7	2ace	0.75	In	Second	Acetylcholinesterase
3.1.1.29	2pth	0.67	In	Best	Peptidyl-trna hydrolase
3.1.1.45	1din	0.62	Out	Best	Dienelactone hydrolase
3.1.1.47	1bwp	0.58	In	Best	Platelet-activating factor acetylhydrolase
3.1.1.61	1chd	1.0	In	Best	Cheb methylesterase
3.1.3.5	1ush	0.89	In	Second	5'-nucleotidase
3.1.3.48	1ytw	0.95	In	Second	Yersinia protein tyrosine phosphatase
3.1.4.3	1ah7	0.55	In	Best	Phospholipase c
3.1.6.8	1auk	1.0	In	Best	Arylsulfatase a
3.1.11.2	1ako	1.0	In	Best	Exonuclease iii
3.1.21.4	1cfr	0.71	In	Second	Restriction endonuclease
3.1.30.1	1ak0	0.93	In	Best	P1 nuclease
3.1.31.1	1a2t	1.0	In	Best	Staphylococcal nuclease
Hydrolases (continued)					
3.2.1.4	2eng	0.78	In	Best	Endoglucanase V
3.2.1.8	1bvv	0.48	In	Second	Endo-1,4-beta-xylanase
3.2.1.8	1exp	1.0	In	Best	Beta-1,4-d-glycanase cex-cd
3.2.1.8	2his	0.76	Out	Second	Cellulomonas fimi family 10 beta-1,4-glycanase
3.2.1.10	1uok	0.83	In	Best	Oligo-1,6-glucosidase
3.2.1.17	206l	0.35	In	Best	Lysozyme
3.2.1.18	1euu	0.61	In	Second	Sialidase
3.2.1.18	7nn9	1.0	In	Best	Neuraminidase N9
3.2.1.21	1cbg	0.83	In	Best	Cyanogenic beta-glucosidase
3.2.1.60	2amg	0.94	In	Best	1,4-alpha-d-glucan maltotetrahydrolase
3.2.1.68	1bf2		Not detected		Isoamylase
3.2.3.1	1myr	0.91	In	Best	Myrosinase
3.4.11.1	1lam	0.71	In	Best	Leucine aminopeptidase
3.4.11.9	1a16	0.8	In	Best	Aminopeptidase p
3.4.17.8	1lbu	0.94	In	Best	Muramoyl-pentapeptide carboxypeptidase
3.4.19.12	1uch	1.0	Out	Best	Ubiquitin c-terminal hydrolase uch-l3
3.4.22.40	1gcb	0.79	In	Second	Gal6 hg (emts) derivative
3.4.23.1	1am5	1.0	In	Second	Pepsin
3.4.24.17	1hfs	0.56	In	Second	Stromelysin-1
3.4.24.17	1slm	0.69	In	Best	Stromelysin-1
3.4.24.36	1lml	0.52	In	Best	Leishmanolysin
3.6.1.3	1kaz	0.63	Out	Best	70kd heat shock cognate protein
3.6.1.7	2acy	1.0	In	Second	Acylphosphatase
3.6.1.29	5fit	0.95	In	Best	Fragile histidine triad protein
3.6.4.6	1nsf	1.0	In	Best	N-ethylmaleimide sensitive factor
3.8.1.5	1b6g	1.0	In	Best	Haloalkane dehalogenase
Lyases : 10					
4.1.1.7	1bfd	0.61	In	Best	Benzoylformate decarboxylase
4.1.1.49	1aq2	0.5	In	Second	Phosphoenolpyruvate carboxykinase
4.1.2.17	1fua	1.0	In	Best	L-fuculose-1-phosphate aldolase
4.1.2.25	2dhn	0.67	In	Best	7,8-dihydroneopterin aldolase
4.2.2.15	1sll	0.79	In	Second	Sialidase 1
4.2.3.9	5eat	1.0	In	Best	5-epi-aristolochene synthase
4.2.99.18	1bix	1.0	In	Best	Ap endonuclease 1
4.2.99.18	2abk	0.51	In	Best	Endonuclease iii
4.3.1.19	1tdj	0.8	In	Best	Biosynthetic threonine deaminase
4.6.1.13	2plc	1.0	In	Best	Phosphatidylinositol-specific phospholipase c

〈표 1〉 계속

Ec num	Pdb code	Representative Precision	Active Site	Cluster Score	Description
Isomerases : 3					
5.3.1.8	1pmi	0.92	In	Best	Phosphomannose isomerase
5.3.1.24	1nsj	0.70	In	Best	Phosphoribosyl anthranilate isomerase
5.99.1.3	1ab4	1.0	In	Best	Gyrase A
Ligases : 7					
6.1.1.10	1a8h	0.92	In	Second	Methionyl-trna synthetase
6.3.2.3	1gsa	0.63	Out	Best	Glutathione synthetase
6.3.2.4	2dln	1.0	In	Best	D-alanine--d-alanine ligase
6.3.2.9	1uag	0.78	In	Best	Udp-n-acetylmuramoyl-l-alanine-:d-glutamate ligase
6.3.3.3	1dae	0.59	In	Best	Dethiobiotin synthetase
6.3.4.4	1gim	0.74	In	Second	Adenylosuccinate synthetase
6.5.1.1	1a0i	Not detected			DNA ligase

였다. 따라서 모든 단백질은 72개 조합으로 실험되었고, 각 단백질에 대한 유효 클러스터는 distance cut-off와 inflation factor의 조합에 따라 한 개 이상이 될 수 있다. 각 단백질의 유효 클러스터 개수는 second cluster score를 갖는 cluster를 고려하면 더 증가된다. 우리는 각 단백질의 대표적인 best cluster를 조사하였고, 대표 클러스터는 정의 3에서 나타내는 바와 같다. 이 과정에서 우리는 낮은 true positive 잔기 개수(< 9)를 갖는 cluster를 제거하였고, 가장 좋은 precision을 갖는 cluster를 선택하였다. 여기서, 한 개 이상의 대표 클러스터를 갖는 단백질도 23개나 됨을 발견하였다.

예를 들어, '1mla'는 best cluster score만 고려한다면 6개 조합을 갖는다. 만약 second cluster score도 고려한다면, 유효 클러스터를 생성하는 조합의 수는 23개로 증가한다. '1mla'는 radius = 12Å와 inflation = 1.7 또는 1.8의 2개 조합으로의 대표 클러스터를 갖고, 두 조합의 precision은 동일하게 1.0이다. <표 2>는 91개 단백질의 대표 클러스터 133개의 distance cut-off value와 inflation factor 빈발 조합을 보여준다. 또한 <표 3>에서는 가장 높은 20개의 유효 클러스터들의 빈발 조합을 나타낸다. 91개 단백질의 대표 클러스터는 낮은 distance cut-off threshold(13Å 근처)와 MCL 알고리즘의 높은 inflation factor(1.7 근처) 일 경우 형성된다. <표 3>의 유효 클러스터의 빈발 조합의 경우도 이와 같은 양상을 나타내었다. 표 1의 'Representative Precision' 열은 대표 cluster의 precision이고, 평균 precision은 '0.83'이다. 또한 대표 클러스터가 활성사이트를 포함하는 경우 <표 1>에서, Active Site 필드에 'In'으로 표기 하였고, 포함하지 않은 경우는 'Out'로 표기하였다.

효소의 활성 사이트는 일반적으로 'cleft' 또는 'cavity'라고 불리는 오목한 영역에 위치한다[21]. 따라서 (그림 9(a))의 어두운 부분처럼, 바인딩 영역도 오목한 영역에 위치한다.

cleft나 cavity를 추출하는 프로그램[10, 22]은 이미 소개되었다. 우리의 실험 결과에서 흥미로운 점은 단백질의 대

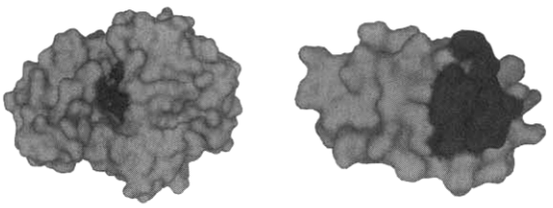
〈표 2〉 대표 클러스터의 빈발 조합

Distance Cut-off	Inflation	No. of representative Clusters
11	1.8	18
11	1.7	13
12	1.7	12
11	1.6	11
12	1.5	9
12	1.6	8
11	1.5	8
12	1.8	6
14	1.7	5
14	1.8	5
11	1.4	5
13	1.7	4
13	1.6	4
16	1.8	4
13	1.8	4
14	1.6	4
17	1.8	4

〈표 3〉 유효 클러스터의 빈발 조합

Distance Cut-off	Inflation	No. of Valid Cluster	Ratio of Valid Cluster
13	1.7	58	64%
14	1.8	58	64%
13	1.8	56	62%
11	1.5	55	60%
12	1.6	54	59%
12	1.7	53	58%
14	1.7	53	58%
11	1.7	50	55%
13	1.6	49	54%
12	1.8	48	53%
15	1.8	48	53%
12	1.5	47	52%
16	1.8	47	52%
11	1.4	46	51%
14	1.6	46	51%
11	1.8	45	49%
15	1.7	42	46%
11	1.6	41	45%
13	1.5	41	45%
17	1.8	40	44%

표 cluster는 (그림 9(b))에서 볼 수 있듯이, cleft나 cavity 이외의 영역에서도 발견이 된다는 점이다. 즉, 우리의 방법은 cavity나 cleft 추출하는 기존의 툴에서 발견할 수 없는 바인딩 잔기를 추출할 수 있다.



(a)1mla (b)2acy

(그림 9) 대표 클러스터의 예

6. 결 론

단백질 기능을 예측하기 위해 단백질의 구조와 서열 분석이 수행되어 왔다. 서열 분석은 서열 상동성을 통해 단백질 기능을 예측한다. 그러나 유사한 폴드 또는 서열을 갖지만 다른 기능을 수행하는 단백질이 밝혀졌다. 그리고 폴드가 유사하지만 활성 사이트가 다르다면 다른 기능을 수행할 수 있다.

이 논문에서 우리는 MCL process와 함께 잔기 간의 connvectivity를 이용하여, 단백질 표면의 활성 사이트 주변에 위치하는 바인딩 영역을 예측하는 방법을 제시하였다. distance matrix에서 의미있는 distance value만을 선택하기 위하여 distance cut-off value를 다양하게 적용하였고, MCL 알고리즘의 다양한 inflation factor를 이용하여 실험하였다. CSA 데이터 (94개 단일 체인 단백질)를 이용한 실험결과, 제시한 방법은 91개 단백질에서 표면의 활성 사이트 주변에 위치하는 바인딩 영역을 추출하였다. 대표 cluster는 평균 83% precision의 결과를 산출하였다. 또한 실험으로 원하는 결과를 가진 클러스터는 비교적 낮은 distance cut-off threshold(13Å 근처)와 MCL 알고리즘의 높은 inflation factor (1.7 근처) 일 경우 형성되었음을 보였다.

이 논문에서 우리는 distance matrix의 잔기 연결 정보를 이용하여 새로운 기하학적 특징을 제시하였다. 우리의 연구는 활성사이트와 관련 없는 잔기는 제거하고 활성 사이트 주변의 잔기를 선택함으로써 활성 사이트 잔기를 예측하는데 이용할 수 있다. 또한, 제시한 방법은 기존의 cavity 또는 cleft 추출 툴은 검출이 불가능한 표면상 볼록한 부분의 바인딩 영역또한 검출 할 수 있는 장점이 있다.

참 고 문 헌

[1] A.Stark, A.Shkumatov, "Finding Functional Sites in Structural Genomics," Proteins. Structure, Vol.12, pp.1205-1412, 2004.

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

[3] P.C Babbitt, "Definition of enzyme function for the structural genomics era," Curr. Opin, Chem. Biol., Vol.7, pp.230-237, 2003.

[4] S. Van Dongen, "Graph clustering by flow simulation," PhD thesis, University of Utrecht, The Netherlands, 2000.

[5] A.J. Enright, S.Van Dongen and C.A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," Nucleic Acids Research, Vol.30, No.7, pp. 1575-1584, 2002.

[6] C.T. Porter, G.J. Bartlett, and J.M. Thornton, "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data," Nucl. Acids. Res., Vol.32, pp.129-133, 2004.

[7] Reichmann D, Rahat O, Cohen M, Neuvirth H, Schreiber G., "The molecular architecture of protein-protein binding sites," Curr Opin Struct Biol., Vol.17, No.1, pp.67-76, 2007.

[8] 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., Vol.303, pp.455-465, 2000.

[9] Fabian Glaser, Richard J. Morris, Rafael J. Najmanovich, Roman A. Laskowski, and Janet M. Thornton, "A Method for Localizing Ligand Binding Pockets in Protein Structures.," PROTEINS: Structure, Function, and Bioinformatics Vol.62, pp.479 - 488, 2006.

[10] R.A. Laskowski, "SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions," J.Mol.Graph, Vol.13, pp.307-308, 1995.

[11] Changhui Yan, Vasant Honavar, and Drena Dobbs, "Predicting Protein-Protein Interaction Sites From Amino Acid Sequence," Department of Computer Science Iowa State University, 2002.

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

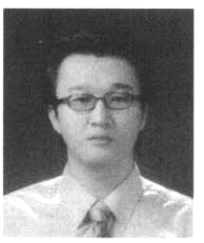
[13] Oliviero Carugo and Giacomo Franzot, "Prediction of protein-protein interactions based on surface patch comparison," Proteomics, Vol.4, pp.1727-1736, 2004.

[14] Susan Jones and Janet M.Thornton, "Principles of protein-protein interactions," Proc. Natl. Acad. Sci. USA, Vol.93, pp.13-20, 1996.

[15] L.Mirny and E.Domany, "Protein Fold Recognition and Dynamics in the Space of Contact Maps," Proteins, pp.391-410, 1996.

[16] M.Vendruscolo, E.Kussell and E.Domany, "Recovery of Protein Structure from Contact Maps," Fold. Des, pp.295-306, 1997.

[17] G. Lancia, R. Carr, B. Walenz, and S.Istrail, "101 optimal PDB structure alignments : a branch-and-cut algorithm for the maximum contact map overlap problem," Proceedings of The Fifth Annual International Conference on Computational Molecular Biology, RECOMB, 2001.
[18] B. Carr, W. E. Hart, N. Krasnogor, E. K. Burke, J. D. Hirst, and J. E. Smith, "Alignment of protein structures with a memetic evolutionary algorithm." In GECCO-2002: Proceedings of the Genetic and evolutionary Computation Conference, Morgan Kaufman, 2002.
[19] E.L.L. Sonnhammer and J.C. Wooton, "Dynamic contact maps of protein structures," Journal of Molecular Graphics and Modelling, Vol.16, pp.1-5, 1998.
[20] http://www.umass.edu/microbio/rasmol/index2.htm
[21] R.A. Laskowski, N.MLuscombe, M.B.Swindells and J.M.Thornton, "Protein clefts in molecular recognition and function," Protein Science, Vol.5, pp. 2438-2452, 1996.
[22] http://xray.bmc.uu.se/usf/voidoo.html.
[23] 김선신, 정광수, 류근호, "단백질 구조기반 단백질 간의 기능관계 예측 기법," 한국정보처리학회, 12권 2호, pp.55-58, 2005.
[24] Kwang Su Jung, Sunshin Kim, Keun Ho Ryu, "A Personalized Biological Data Management System based on BSML," LNBI, vol.4115, pp.362-371, 2006.
[25] Sunshin Kim, Kwang Su Jung, Keun Ho Ryu, "Automatic Orthologous-Protein-Clustering from Multiple Complete-Genomes by the Best Reciprocal BLAST Hits," LNBI, Vol.3916, pp.60-70, 2006.
[26] 유기진, 정광수, 류근호, "잔기 위치 예측을 위한 단백질 기하학적 특징 추출 기법," 한국정보처리학회, 13권 2호, pp.673-676, 2006.
[27] 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, 2007.
[28] 유기진, 정광수, 류근호, 정용제, "단백질 활성 사이트 비교를 통한 단백질 기능 예측 기법 설계," 한국데이터베이스학회, pp.191-197, 2007.
[29] 김선신, 이충세, 류근호, "유전체 상호간의 BLAST 최대 히트를 사용하여 서열화가 완성된 다수의 유전체로부터 Orthologous 단백질그룹을 자동적으로 클러스터링하는 기법," 한국정보처리학회논문지D, 13D권 2호, pp.207-214, 2006.
[30] 김선신, 이범주, 정광수, 김영창, 김태경, 조완섭, 이충세, 류근호, "다종의 유전체로부터 탐지된 올소로그(Ortholog)군집에 대한 분석," 한국정보과학회논문지 출판예정, 2008.

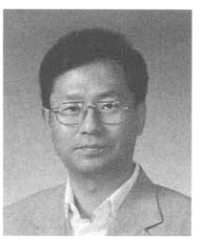


정 광 수

e-mail : ksjung@dblab.chungbuk.ac.kr
2001년 충북대학교 화학공학부(공학사)
2004년 충북대학교 정보산업공학과 (공학석사)
2006년 충북대학교 대학원 전자계산학과 박사수료
관심분야 : Bioinformatics, 단백질 서열 및 구조, 생명정보 데이터베이스, 데이터마이닝

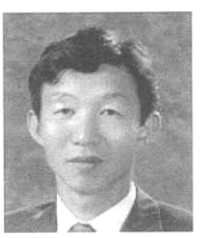
유 기 진

e-mail : heyu4580@dblab.chungbuk.ac.kr
2006년 충북대학교 생물학과(학사)
2006년 충북대학교 대학원 전자계산학과 입학
2006년 ~ 현 재 충북대학교 대학원 전자계산학과 석사과정
관심분야 : 바이오인포메틱스, 단백질 구조, 데이터마이닝



정 용 제

e-mail : chungyj@chungbuk.ac.kr
1979년 연세대학교 이과대학 화학과 (학사)
1981년 서울대학교 대학원 화학과 (이학석사)
1989년 미국 피츠버그대학교 결정학과 (이학박사)
1989년 ~ 1991년 피츠버그대학교 박사후연구원
1991년 ~ 현 재 충북대학교 생명과학부 교수
관심분야 : X-선 결정학, 단백질표면구조 분석 및 예측 등



류 근 호

email : khryu@dblab.chungbuk.ac.kr
1976년 숭실대학교 전산학과(이학사)
1980년 연세대학교 공업 대학원 전산전공(공학석사)
1988년 연세대학교 대학원 전산전공(공학박사)
1976년 ~ 1986년 육군 군수 지원사 전산실(ROTC 장교), 한국전자통신연구소(연구원), 한국 방송대학교 전산학과(조교수) 근무
1989년 ~ 1991년 University of Arizona, Research Staff(TempIS 연구원, Temporal DB)
1986년 ~ 현 재 충북대학교 전기전자컴퓨터 공학부 교수
관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 지식기반 정보검색 시스템, 유비쿼터스컴퓨팅 및 스트림데이터처리, 데이터마이닝, 데이터베이스 보안, 바이오인포메틱스 등