Motif 기반의 단백질 군집화

진 훈[0] 김현식 김인철
경기대학교 전자계산학과
{jinun, advance7, kic}@kyonggi.ac.kr

Motif-Based Protein Clustering

Hoon Jin[0] Hyun-Sik Kim, In-Cheol Kim
Department of Computer Science, Kyonggi University

요 약

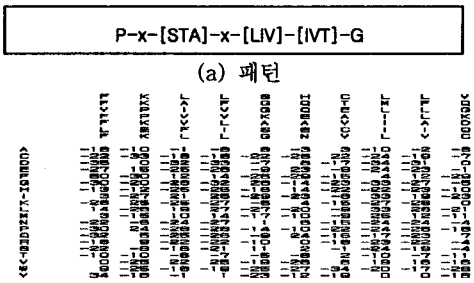
motif란 기능적으로 유사한 단백질 군의 아미노산 서열들에 공통적으로 나타나는 일정한 패턴이나 부분서열을 말한다. 본 논문에서는 motif들로 각 단백질의 특성을 표현한 다음, 이것을 기초로 유사성을 비교하여 단백질들을 기능적으로 유사한 여러 개의 계층적 군으로 나누는 군집화 방법을 소개하였다. 영역 특성상 확장성과 계층성을 가지는 신경망 GHSOM을 군집화 알고리즘으로 사용하였고, 실제 307 개의 단백질들에 대한 군집화 실험을 통해 그 효과를 확인해보았다.

1. 서론

휴먼 게놈 프로젝트 이후 최근 생명과학분야의 주된 관심사는 그 동안 찾아진 미지의 유전자나 그것이 발현되어 만들어지는 단백질의 기능과 구조가 과연 무엇인가를 밝히는 문제에 집중되고 있다. 특히 미지의 단백질 서열데이터로부터 그것의 구조와 기능을 예측하는 일은 대단히 어려운 작업으로서, 주로 서열정렬(sequence alignment)을 통해 서열이 비슷하면서 이미 잘 알려진 다른 단백질들과 연관시켜 이 문제를 해결하려는 시도가 많이 이루어지고 있다. 본 논문에서는 단순히 서열데이터를 바로 비교하기보다는 각 단백질 서열에 포함된 motif들을 특징으로 삼아 가능한 많은 motif를 공유하는 단백질들끼리 서로 묶음으로써 구조적으로나 기능적으로 유사한 단백질 군(protein family)으로 군집화(clustering)하려고 하였다. 이를 위해 대표적인 motif 데이터베이스인 Prosite와 motif 분석프로그램인 ScanProsite를 이용하여 군집화를 위한 데이터를 생성하였고, 확장성과 계층성을 가진 군집화 알고리즘인 신경망 GHSOM을 적용하여 군집화를 시도하였다.

2. Motif와 단백질 군

일반적으로 motif란 기능적으로 유사한 단백질 군의 아미노산 서열(amino acid sequence)들에 공통적으로 나타나는 일정한 패턴이나 부분서열을 말한다. 이러한 단백질 서열의 특정 영역은 단백질의 서열에서뿐만 아니라 구조(structure)에서도 잘 보존되는 것으로 알려져 있어, 한 단백질의 2차 혹은 3차원 구조를 예측하거나 그 단백질의 효소적 작용이나 그 밖의 성질을 판단하는데 중요한 역할을 한다. 미지의 한 단백질의 구조와 기능을 예측하기 위한 가장 보편적인 방법으로는, 이미 구조나 기능이 알려져 있는 단백질들과 서열 짝 정렬(pairwise alignment)을 통해 서열의 유사성을 알아보는 방법을 많이 이용한다. 하지만 이러한 짝 정렬을 통해 전체 서열의 유사성이 낮아보이는 두 단백질의 경우라도 특정 motif를 공유하고 있으면 구조나 기능 면에서 밀접한 관련이 있다고 판단한다.



(a) 패턴

(b) 프로필

그림 1 motif 표현방식

일반적으로 하나의 motif는 (그림 1)의 (a)와 같은 정규식 형태의 패턴(pattern)으로 표현하거나, 좀더 정량적 혹은 정성적으로 표현하기 위해 (b)와 같은 프로필(profile)로 표현한다. 한편, 프로필은 (그림 1)의 (b)와 같은 점수행렬이외에도 은닉 마르코프 모델(HMM) 형태로 표현하기도 한다. Prosite, Pfam, Prints 등은 motif를 중심으로 단백질 군을 분류해 놓은 대표적인 motif 데이터베이스들이며, ScanProsite, MotifScan 등은 이러한 데이터베이스들을 참조하여 한 단백질 서열에 포함된 motif들을 찾아주는 서비스 프로그램들이다.

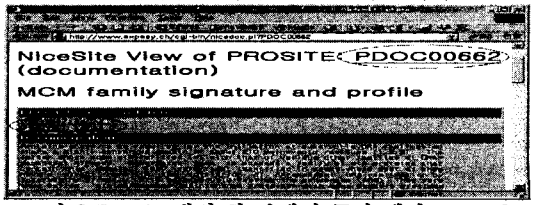


그림 2 Prosite내의 한 단백질 군과 해당 motif들

(그림 2)는 Prosite 데이터베이스가 포함하고 있는

약 1100여 개의 단백질 군중의 하나인 PDOC0062에 대해 기술되어 있는 내용으로서, 이 단백질 군은 DNA 복제 초기화에 필요한 기능을 가지고 있다는 설명과 더불어 이 단백질 군을 특징짓는 2개의 motif인 PS00847과 PS50051을 나타내고 있다. 이와 같이 대부분의 기존 motif 데이터베이스들은 하나 또는 많아야 두 개의 특징적인 motif들과 하나의 독특한 기능을 나타내는 단백질 군을 서로 연결해놓고 있다. 즉, 대부분이 하나의 motif에 대해 하나의 단백질 군과 하나의 기능을 결부시켜 놓았다. 하지만 많은 경우, 하나의 단백질은 여러 개의 서로 다른 motif들을 포함할 수 있으며, 또한 하나의 단백질은 조건과 상황에 따라 여러 가지 기능을 가질 수 있는 것으로 알려져 있다. 실제로도 하나의 단백질이 어떤 motif들을 포함하고 있느냐에 따라 Prosite 데이터베이스의 분류 체계상 여러 단백질 군에 속하는 경우를 자주 확인할 수 있다.

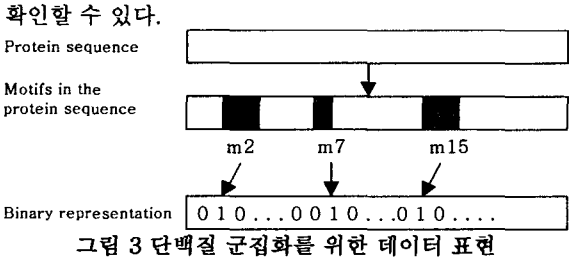


그림 3 단백질 군집화를 위한 데이터 표현

3. 단백질 군집화 문제

motif에 기초하여 유사 단백질 군을 찾고자 하는 대표적인 연구의 하나인 Wang[2]의 연구에서는 각 motif와 해당 단백질 군을 정의해 둔 motif 데이터베이스를 훈련 데이터 집합(training dataset)으로 삼아 일반적인 분류지식을 이끌어낸 뒤, 이것을 이용하여 미지의 한 단백질을 motif 데이터베이스에 정의된 하나의 단백질 군으로 분류하였다. 하지만 앞서 설명한 바와 같이 motif 데이터베이스에서 규정하고 있는 motif와 단백질 군과의 대응관계는 대부분이 1:1의 관계로서, 분류지식을 나타내는 결정트리(decision tree)는 매우 긴 하나의 경사트리(skew tree)를 이루게 된다. 그리고 이 트리의 각 비 단말노드(non-leaf node)는 가능한 모든 motif 중의 하나가 되고 각 단말노드(leaf node)는 그 motif에 대응되는 하나의 단백질 군을 가리킨다. 즉, 주어진 한 단백질의 유사 단백질 군을 찾는 문제를 훈련집합에 기초한 하나의 자동 분류(classification) 문제로 정의한 접근법은 단순히 그 단백질이 가진 단 하나의 특정 motif에 기초하여 하나의 기능 단백질 군으로만 분류하려고 함으로써 원래 motif 데이터베이스가 제공하는 정보이외에 더 이상의 특별한 단백질 간의 연관지식을 제공하지 못하는 것으로 파악된다.

본 연구에서는 한 단백질의 유사 단백질 군을 찾는 문제를 하나의 분류(classification) 문제로 보지 않고, 단백질들이 가진 다양한 motif들에 기초하여 유사한 단백질들을 묶는 하나의 군집화(clustering) 문제로 파악한다. 다시 말해 본 연구에서는 하나의 motif는 곧 그 단백질이 가지는 것으로 추정되는 하나의 기능적 특징으로 보고, 각 단백질에 포함된 모든 motif, 즉 모든 기능적 특징들을 함께 고려하여 유사성을 계산하고 단백질들을 군집화 하려고 한다. 이를 위하여 군집화 대상이 되는 각 단백질들은 (그림 3)과 같은 전처리 과정(preprocessing)을 거친다. 먼저 ScanProsite나 MotifScan등과 같은 프로그램을 이용하여 해당 단백질에 포함된 motif들을 찾아낸다. 다음은 이 motif들을 특징집합(feature set)으로 삼아 이진 벡터형태의 데이터로 변환한다. 이 이진 벡터에서는 특정 motif가 그 단백질에 포함되어 있으면 1, 그렇지 않으면 0으로 표현된다. 끝으로, 이렇게 이진 벡터형태로 표현된 단백질 데이터들을 대상으로 하나의 군집화 알고리즘을 적용함으로써 몇 개의 유사 단백질 군으로 나눈다. 미지의 단백질에 대해서도 이와 같은 군집화 과정을 통해 유사 단백질 군을 찾을 수 있다.

이러한 단백질 군집화를 위해서는 다양한 알고리즘들을 적용할 수 있다. 하지만 영역 특성에 따른 다음과 같은 몇 가지 요구사항을 만족할 수 있는 군집화 알고리즘이 바람직하다. 첫째는 확장성이다. 일반적으로 단백질 서열상의 motif수는 현재 밝혀진 것들 외에 추가적으로 드러날 수 있는 새로운 motif의 수가 많지 않으리라고 예상하는 반면, 인간과 같은 고등동물을 비롯해 한 생명체를 이루는 단백질의 종류는 그 수를 헤아리기 어려울 정도로 많고 그 구조나 기능도 다양하다. 따라서 새로운 단백질들은 새로운 군집을 형성할 수 있으며, 기존의 어떤 단백질들을 가지고 군집화를 하였느냐에 따라 군집의 수, 모양, 그리고 밀도가 크게 달라질 수 있다. 따라서 단백질 군집화에 적용될 알고리즘은 군집의 개수를 미리 정해주지 않고, 유입되는 단백질들의 유형별 분포에 따라 유연하게 확장 가능하여야 한다. 요구되는 두 번째 성질은 계층성이다. 많은 수의 단백질들을 대상으로 motif에 기초한 기능적 유사성에 따라 체계적인 분류 계통도를 이끌어내기 위해서는 그들에 대한 수평 분할적 군집화보다는 계층적 군집화가 요구된다.

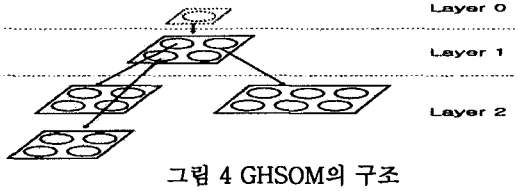


그림 4 GHSOM의 구조

4. GHSOM을 이용한 군집화

GHSOM (Growing Hierarchical SOM)은 성장 SOM(Growing SOM)과 계층적 SOM(Hierarchical SOM)의 장점을 결합하여 만들어진 군집화 알고리즘이다. GHSOM의 기본적인 구조는 (그림 4)과 같이 여러 계층의 서로 독립적인 SOM들로 구성된 계층적 구조로 되어있고, 기존 SOM과는 달리 입력 데이터들에 따라 맵(map)의 크기와 계층 수가 스스로 늘어나는 성질을 가지고 있다. GHSOM에서 계층은 크게 계층 0, 계층 1, 그리고 나머지 부분으로 구분 할 수 있다. 우선 계층 0은 가상의 계층으로 1개의 유닛(unit)을 포함하고 있으며, 이 유닛의 가중치벡터는 $m_0 = [\mu_{01}, \mu_{02}, \ldots, \mu_{0n}]^T$ 와 같이 표현되며, 모든 입력 데이터의 평균으로 초기화 된다.

입력데이터 x 와 이 유닛과 편차는 (식 1)과 같이 나타낼 수 있다.

$$mqe_0 = \frac{1}{d} \cdot \|m_0 - x\| \quad (식\ 1)$$

, d: x의 수 (식 1)

mqe_0를 계산한 후에 GHSOM의 첫번째 SOM으로부터 훈련이 시작된다. 첫번째 계층의 맵은 유닛의 수보다 적은 수의 유닛으로 초기화된다. 각 유닛 i는 (식 2) 같이 n-차원의 벡터 m_i로 정의된다.

$$m_i = [\mu_{i1}, \mu_{i2}, \ldots, \mu_{in}]^T, m_i = \Re^n \quad (식\ 2)$$

각 유닛들은 랜덤한 값으로 초기화되며, GHSOM의 학습 규칙은 (식 3)과 같다. 여기서 a 는 학습률이고, h_{ci}는 이웃함수(neighborhood function)이고, x는 현재의 입력 패턴이다. 그리고 c는 t 만큼 반복한 후의 승자 유닛이다.

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)] \quad (식\ 3)$$

일정한 회수만큼의 훈련이 이루어진 후에 (식 4)에 의해 맵의 MQE(Mean Quantization Error)가 계산된다. 여기서 u는 SOM m 에 포함된 유닛 i의 갯수이고, mqe_i에

$$MQE_m = \frac{1}{u} \cdot \sum_i mqe_i \quad (식\ 4)$$

의해서 계산된다.

그리고 (식 5)의 조건이 만족하는 동안 mqe_e가 가장 큰 유닛 e 에 새로운 열이나 행을 삽입함으로써 맵은 계속 성장한다.

$$MQE_m \geq \tau_m \cdot mqe_0 \quad (식\ 5)$$

일단 한 계층의 성장이 종료되면, 이 맵은 다음 계층으로 확장을 시도한다. 이때 매우 높은 mqe를 가진 이 유닛들은 다음 계층의 새로운 맵에 추가된다. 그리고 각 유닛 i는 (식 6)과 같은 조건을 만족하면 확장하게 된다.

$$mqe_i > \tau_u \cdot mqe_0 \quad (식\ 6)$$

따라서 이러한 특징을 가진 GHSOM은 앞서 설명한 단백질 군집화 알고리즘의 요구 성질들을 잘 만족하는 것으로 판단된다.

5. 군집화 실험

본 연구에서는 단백질 서열 데이터베이스인 SWISSPROT로부터 임의로 선택한 단백질 307개에 대하여 앞서 설명한 방식대로 신경망 GHSOM을 이용하여 군집화를 시행하였다. 먼저 ScanProsite를 실행하여 각 단백질 서열에 포함된 Prosite 데이터베이스상의 motif들을 검출하였다. 총 307개의 단백질 서열에 포함된 서로 다른 64개의 motif들을 특징집합으로 정하고, 이를 바탕으로 이진벡터 형태의 데이터들을 생성하였다. 가능한 모든 단백질에 대한 군집화를 고려한다면 Prosite 데이터베이스에 포함된 약 1300개의 motif 전부를 단백질을 표현하는 특징집합으로 사용하여야 할 것이나 실험의 효율을 위해 본 실험의 대상이 되는 단백질 집합에 실제 등장하지 않는 motif들을 제외한 총 64개의 motif들만을 사용하여 각 단백질을 표현하였다. 군집화를 위한 신경망 GHSOM의 학습을 위해서는MATLAB으로 구현한 GHSOM Toolbox를 이용하였다. GHSOM의 파라미터들인 τ_m 과 τ_u는 몇 번의 실험을 통해 각각 0.5와 0.01로 설정하였다. (그림 5)는 이와 같은 GHSOM을 이용한 군집화를 통해 얻어진 단백질들의 군을 나타내고 있다. 그림에서 각 셀(cel)은 신경망 GHSOM의 하나의 유닛에 해당되며 이는 곧 하나의 단백질 군을 나타낸다. 검은 색 바탕의 셀들인 제 1 계층은 총 72 개의 군집들로 이루어져 있고, 흰색 바탕의 셀들인 제 2 계층은 가장 세분화된 단백질 군들로서 총 86개의 군집들로 이루어져 있다. 각 셀들 위에는 그 단백질 군에 속하는 단백질들의 이름이 표시되어 있다. 그림에서 타원으로 표시된 가장 깊은 레벨의 동일 군집에 속한 단백질 Q05670과 Q09763은 문헌자료를 통해 조사해본 결과 두 단백질 모두 Dbl 상동성(DH) 도메인에 속한 단백질들로서 유사한 세포작용을 하는 것으로 밝혀졌다. 이들과 멀리 떨어진 군집에 속한 또 다른 단백질 P48562는 이 두 단백질들과는 달리 pleckstrin 상동성(PH) 도메인에 속한 단백질로서 전혀 다른 기능을 가진 것으로 밝혀졌다. 그 밖의 대부분 단백질 군들도 Prosite 데이터베이스상 분류체계와 모순을 보이지 않았다. 하지만 아직 큰 숙제로 남아있는 이와 같은 다 기능적 단백질 군집에 대한 생물학적 검증은 매우 어려운 작업이 될 것으로 보인다.

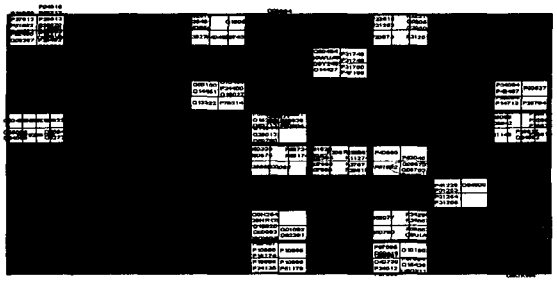


그림 5 GHSOM에 의해 구해진 계층적 단백질 군

6. 결론

본 연구에서는 내포하고 있는 motif들로 각 단백질의 특성을 표현한 다음, 이것을 기초로 유사성을 비교하여 단백질들을 기능적으로 유사한 여러 개의 계층적 군으로 나누는 군집화 방법을 소개하였다. 영역 특성상 확장성과 계층성을 가지는 신경망 GHSOM을 군집화 알고리즘으로 사용하였고, 실제 307 개의 단백질들에 대한 군집화 실험을 통해 그 효과를 확인해보았다. 향후연구로는 군집화 결과에 대한 충분한 후속 검증작업이 필요할 것으로 판단된다.

참고 문헌

[1] L. Falqet, et al, "The PROSITE database, its status in 2002", Nucleic Acids Research, vol.30, No.1, pp.235-238, 2002.

[2] D. Wang, X. Wang, V. Honavar, "Data-Driven Generation of Decision Trees for Motif-Based Assignment of Protein Sequences to Functional families", Proceedings of the Atlantic Symposium on Computational Biology, 2001.

[3] M. Dittenbash, D. Merkl, A. Rauber, "The Growing Hierarchical Self Organizing Map", Proceedings of IJCNN-2000, pp.15-19, 2000.