JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 20, No. 9 : 1816~1821 Sep. 2016

단백질 서열의 상동 관계를 가중 조합한 단백질 이차 구조 예측

지상문*

Prediction of Protein Secondary Structure Using the Weighted Combination of Homology Information of Protein Sequences

Sang-mun Chi*

Department of Computer Science and Engineering, Kyungsung University, Busan 48434, Korea

요 약

단백질은 대부분의 생물학적 과정에서 중대한 역할을 수행하고 있으므로, 단백질 진화, 구조와 기능을 알아내기 위하여 많은 연구가 수행되고 있는데, 단백질의 이차 구조는 이러한 연구의 중요한 기본적 정보이다. 본 연구는 대규모 단백질 구조 자료로부터 단백질 이차 구조 정보를 효과적으로 추출하여 미지의 단백질 서열이 가지는 이차 구조를 예측하려 한다. 질의 서열과 상동관계에 있는 단백질 구조자료내의 서열들을 광범위하게 찾아내기 위하여, 탐색에 사용하는 프로파일의 구성에 질의 서열과 유사한 서열들을 사용하고 갭을 허용하여 반복적인 탐색이 가능한 PSI-BLAST를 사용하였다. 상동 단백질들의 이차구조는 질의 서열과의 상동 관계의 강도에 따라 가중되어 이차 구조 예측에 기여되었다. 이차 구조를 각각 세 개와 여덟 개로 분류하는 예측 실험에서 상동 서열들과 신경망을 동시에 사용하여 93.28%와 88.79%의 정확도를 얻어서 기존 방법보다 성능이 향상되었다.

ABSTRACT

Protein secondary structure is important for the study of protein evolution, structure and function of proteins which play crucial roles in most of biological processes. This paper try to effectively extract protein secondary structure information from the large protein structure database in order to predict the protein secondary structure of a query protein sequence. To find more remote homologous sequences of a query sequence in the protein database, we used PSI-BLAST which can perform gapped iterative searches and use profiles consisting of homologous protein sequences of a query protein. The secondary structures of the homologous sequences are weighed combined to the secondary structure prediction according to their relative degree of similarity to the query sequence. When homologous sequences with a neural network predictor were used, the accuracies were higher than those of current state-of-art techniques, achieving a Q3 accuracy of 92.28% and a Q8 accuracy of 88.79%.

키워드 : 단백질 이차 구조, 상동 단백질 탐색, PSI-BLAST, 서열 유사성

Key word : Protein secondary structure, Homologous protein search, PSI-BLAST, Sequence similarity

Received 30 May 2016, **Revised** 01 June 2016, **Accepted** 17 June 2016

* **Corresponding Author** Sang-Mun Chi (E-mail:smchiks@ks.ac.kr,Tel:+82-51-663-5146)
Department of Computer Science and Engineering, Kyungsung University, Busan 48434, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2016.20.9.1816 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

단백질은 생명현상에 필요한 대부분의 생화학 반응을 수행하므로, 그 기능과 구조에 대한 많은 연구가 수행되어 왔다. 컴퓨터 기술의 발전에 따라 단백질을 구성하는 아미노산 서열 정보만으로 단백질의 구조, 기능과 진화 과정을 예측하려는 시도가 활발하다. 이러한 예측 방법들을 뒷받침하는 중요한 구성 요소 기술로서 단백질의 이차 구조 예측이 사용된다. 단백질의 일차 구조는 단백질을 구성하는 아미노산 서열을 의미하고, 이차 구조는 이들 아미노산의 카보닐기 산소 원자와 아민기 수소 원자 사이에 수소 결합이 형성되어 열역학적으로 안정하기 때문에 나타나는 국소적으로 규칙적 구조인 나선이나 병풍 등의 구조적 형태를 의미한다[1].

현재의 단백질 이차 구조 예측과 단백질 구조 예측 방법들은 입력으로 단백질 서열을 직접적으로 사용하지 않고, 입력 서열과 유사한 서열들을 단백질 자료에서 찾아내어 이들의 평균적인 단백질 서열을 사용한다[2-5]. 이것은 질의 서열만을 사용하는 것보다 유사 서열들로 이루어지는 단백질 패밀리에서 이차 구조 등의 단백질 구조가 더 잘 보존되므로, 단백질 프로파일을 사용하는 것이 서열을 직접 사용하는 것보다 예측 성능이 높기 때문이다. 이밖에도 단백질 이차 구조 예측에 진화 정보를 이용하는 방법은 입력 서열과 유사한 서열들의 예측 결과를 각각 구하여 이들 결과의 투표로서 최종 예측을 결정하는 방법이 있다[3]. 또한, 최근에는 딥러닝 방법이 단백질 이차 구조 예측에도 적용되어 성능 향상이 이루어지고 있다[6-9].

단백질 구조에 대한 분석 기술이 발전함에 따라서 단백질의 삼차원 구조가 밝혀진 단백질 자료의 크기가 매우 커지고 있다. 따라서 단백질 자료로부터 질의 단백질 서열과 상동 관계가 높은 단백질을 찾을 수 있는 가능성이 점점 커지고 있다. 1996년부터 2013년까지 PDB (Protein Data Bank)에 추가되는 단백질 자료를 조사해 보면, 새롭게 추가되는 단백질 서열은 서열 유사성이 30%이상인 것이 기존의 자료에 이미 존재하는 비율이 증가하고 있으며, 2013년에는 94%이상이었다[3]. 본 논문에서는 단백질 구조 자료에 존재하는 질의 서열과 유사한 단백질의 이차 구조를 사용하여 이차구조를 예측한다. 이는 모든 단백질이 가질 수 있는 구조의 형태가 대부분 알려지고 있으므로, 이차 구조를 예측하고자 하는 단백질 서열과 부분적으로 매우 유사한 영역들이 존재할 가능성이 점점 커지고 있기 때문이다. 본 논문에서는 이들 유사한 부분 영역들을 단백질 구조 자료에서 탐색하고, 이들 각 영역과 질의 서열과의 상동 관계의 정도를 효과적으로 최종 분류에 반영하여 이차 구조를 예측하려 한다.

Ⅱ. 단백질 서열의 상동 관계 정보를 조합한 이차구조 예측

2.1. 선행 연구

대부분의 이차 구조 예측 방법들은 단백질 서열의 진화정보를 이용하기 위하여, 이차 구조를 예측하고자 하는 단백질 서열과 유사한 서열들의 정보를 활용한다. 예를 들어, PSIPRED[2]는 이차 구조를 예측하기 위하여 PSI-BLAST(Position Specific Iteration-Basic Local Alignment Search Tool)[10]로 만든 PSSM (position-specific scoring matrix)을 입력으로 사용한다. PSSM은 N x 20의 원소 개수를 가지고, N은 나타내고자 하는 단백질의 부분 영역에 해당하는 서열의 길이이고, 20은 아미노산의 종류이다. 이것은 부분 서열의 각 위치에서 20개의 아미노산과 치환 빈도를 표시하며, 주어진 아미노산 서열과 성질이 비슷한 서열을 찾거나 서로 비교할 때 서열을 대신하여 사용된다. 각 위치별 특성이 반영된 치환빈도를 표시하는 행렬이 국소적인 구조의 특성을 잘 반영하고 있을 것이므로 이를 활용한 예측 및 비교의 정확도는 단순 서열을 사용하였을 때 보다 우수하다고 알려져 있다. PSIPRED는 PSSM에서 가운데 아미노산의 이차 구조를 이 PSSM의 이차 구조로 결정하여 신경망을 학습하는 자료로 사용한다.

단백질 서열의 진화적인 정보를 이용하는 또 다른 방법인 SSpro[3]의 경우에는 세단계로 구성되어 있다. 첫 단계는 PSI-BLAST를 UNIREF50 단백질 자료[11]를 대상으로 세 번 반복하여 다중 서열 정렬을 수행하고, 이 결과를 이용하여 단백질 서열의 각 위치에서 나타나는 아미노산의 빈도를 구한다. 두 번째 단계에서는 첫 단계에서 얻은 단백질 프로파일을 신경망의 입력으로 사용하여 단백질 이차구조를 예측한다. 세 번째 단계에서는 대규모 단백질 구조 자료인 PDB[12]에서 이차 구조를 예측하고자 하는 질의 단백질 서열의 부분서열들

과 유사한 부분 서열을 찾는다. 유사한 부분 서열은 최소한 10개 이상의 연속된 아미노산이 갭이 없이 질의 서열과 일치하고, 10^{-9}이하의 BLAST 기댓값을 갖고, 45%이상의 아미노산이 일치하고, 55%이상의 양의 치환값을 가져야 유사한 서열로 판정하는 조건을 가진다. 질의 서열의 각 위치에 대응하는 유사 부분 서열의 이차 구조를 수집하여, 가장 빈도가 높은 이차 구조로 그 위치의 이차 구조로 예측한다. 두 번째 단계에서 얻은 예측 결과는 세 번째 단계에서 예측되는 결과가 없는 위치에 사용한다. 이밖에 DISTILL Porter[4] 방법은 SSpro의 첫 단계와 두 번째 단계는 유사하고 세 번째 단계만 다른데, 신경망을 사용하여 두 번째 단계의 결과를 필터링하여 최종 결과를 얻는다.

2.2. 제안 방법

본 연구에서는 대규모 단백질 삼차원 구조 자료인 PDB[12]를 보다 효과적으로 이용하여 이차 구조를 예측하려 한다. PDB에 존재하는 유사서열의 삼차원 구조에서 이차 구조를 추출하는 과정에서, 기존 방법은 갭을 허용하지 않고 정확히 일치하는 부분 서열의 정보만을 이용하였지만, 본 연구에서는 질의 서열과 유사한 서열을 보다 광범위하게 탐색한다. 이를 위하여 (1) 탐색에 갭을 허용하고, (2) PSI-BLAST에서 매 반복마다 프로파일의 구성에 추가할 유사서열을 결정하는 *e*값 문턱치를 결정하는 옵션 *h*의 다양한 값을 조사하여 최적값을 알아보고, (3) 이차 구조 예측에 사용하는 서열의 *e*값 최적 문턱치를 조사하고, (4) 이러한 탐색 결과로 얻은 서열의 상동 관계의 강도를 반영하여 조합하였다.

(1) 탐색에 갭 허용

갭을 허용하지 않을 때보다 갭을 허용하여 탐색하면 질의 서열과 유사한 서열을 더 많이 발견할 수 있다. 하지만, 매칭 되지 않는 부분의 서열을 처리하여야 한다. 본 논문에서는 갭에 해당하는 부분의 이차 구조는 예측하지 않는 방법을 사용하였다.

```
Query: 280 GI IGP - - PQEMQPETLQKK 296
           G +G       + +  P       K
Sbjct: 149 GELGSSSNWQLDPM- - AGK 165
```

Fig. 1 Example of a sequence alignment

위의 그림 1은 PSI-BLAST내의 하나의 프로그램인 blastpgp를 사용하여 다중 서열 정합을 수행한 결과들의 일부분이다. 첫 행은 질의 서열의 280번째부터 296번째까지의 부분 아미노산 서열을 나타내고, 세 번째 행은 질의 서열과 유사한 데이터베이스내의 부분서열의 149번째부터 165번째까지를 표시한다. 두 번째 행은 각 위치에서의 정합의 정도를 나타내는데, 같은 아미노산의 경우에는 그 아미노산 문자를 적고, 유사한 아미노산이 대응될 경우에는 +로 표시한다. 첫 번째 행의 질의 서열과 세 번째 행의 유사 서열 모두에 갭이 없는 위치에서만 이차구조 정보를 사용하였다. 즉, 질의서열의 280-284 위치의 이차 구조의 예측을 위해서는 대응되는 유사 서열의 149-153을 이용한다. 하지만, 유사서열의 위치, 154와 155처럼 대응되는 질의 서열이 없거나, 질의 서열 위치 292와 293처럼 대응되는 유사 서열이 없는 경우는 이차 구조정보를 수집하지 않았다.

(2) 프로파일 구성에 사용하는 서열의 *e*값

본 논문에서는 blastpgp를 서열 탐색에 사용한다. 이 방법은 우선 질의 서열과 유사한 단백질 서열들을 데이터베이스에 찾아서 프로파일인 PSSM을 구성한다. 이후에는 질의서열 대신에 프로파일을 이용하여 유사한 서열들을 찾고, 이를 이용하여 다시 프로파일을 구성하는 과정을 반복한다. 새로 만들어질 프로파일에 추가할 서열을 결정하기 위하여, 단백질 서열 자료 중에서 질의 서열과의 *e*값이 옵션 *h*로 주어지는 문턱치보다 낮을 때만 추가한다. 여기서, *e*값은 두 서열이 우연히 일치하는 정도로서, 작은 값일수록 유사성이 높다. 따라서 높은 문턱치 옵션 *h*를 사용하면 더 많은 유사 서열을 사용하여 프로파일이 만들어지므로 질의 서열의 많은 부분과 대응되는 단백질 자료를 찾을 수 있다. 하지만, 반복적으로 프로파일을 만드는 과정 중에 질의 서열과 일치하지 않는 자료들이 추가됨으로서, 초기 질의 서열과는 특성이 다른 프로파일로 표류하는 단점이 있다. 본 논문에서는 옵션 *h*를 조절하여 이차 구조 예측에 최적인 값을 조사한다.

(3) 이차 구조 예측에 사용하는 서열의 *e*값

탐색된 상동성이 있는 단백질 서열 자료 중에서 질의 서열과의 *e*값이 정해진 문턱치 이하인 것만을 사용한다. 이러한 문턱치가 크면 더 많은 유사 서열의 정보를

이차 구조의 예측에 사용하므로 많은 질의 서열의 위치에서 예측이 가능하나, 질의 서열과 유사성이 떨어지는 정보를 사용하는 단점도 존재하여 예측의 정확도를 감소시킬 수 있다.

(4) 유사한 서열들의 가중 조합

단백질 서열 자료를 대상으로 질의 서열과 유사한 서열들을 탐색하면, 그림 1같은 서열간의 정합이 여러 개 나타난다. 따라서 질의 서열의 하나의 위치에 대응되는 여러 개의 서열이 나타나고, 이들의 이차 구조를 조합하여 예측을 수행하였다. 본 논문에서는 질의 서열과 정합된 서열에서 양의 값으로 치환된 개수가 클수록 큰 가중치를 가지도록 조합하였다. 그림 1의 경우에 두 번째 행에서 공백이 아닌 위치의 개수가 양의 값으로 치환된 개수이다. 질의 서열의 하나의 위치에 대응되는 n개의 부분 서열의 위치에서의 각각의 이차 구조를 $s_1, s_2, \cdots, s_n$, 각 n개의 부분 서열이 질의 서열과의 정합될 때 양의 값으로 치환된 백분율을 $p_1, p_2, \cdots, p_n$이라고 하자. 이 위치에서 이차 구조가 m인 점수를 다음 식으로 정의하였다.

$$score_m = \sum_{i=1}^{n} \delta(s_i, m) w^{p_i} \quad (1)$$

단, $\delta(s_i, m)$은 s_i와 m이 같을 때 1이고, 아니면 0이다. w는 실험에 의해서 결정할 파라미터이고, 양의 값으로 치환된 개수가 많을수록 커다란 가중치를 갖도록 w를 p_i번 곱하여 주었다. 이러한 이차 구조의 점수에서 가장 큰 점수를 갖는 이차 구조를 그 위치의 이차 구조로 예측하였다.

Ⅲ. 실험 및 결과

실험에 사용된 자료[3]는 2013년 8월 20일 까지 PDB[12]에 수집된 자료 중에서 해상도가 2.5 옹스트롬 이하이고, 체인이 연속되어 있고, 5개 이하의 미지 아미노산을 가지고, 서열의 길이가 30개 이상이며, 25%이하의 서열 동일성을 가진 5772개의 단백질 자료이며, 이차 구조를 예측하여야할 아미노산의 위치는 1,031,455개이다.

아미노산 서열로부터 이차 구조를 예측하는 방법들을 학습하거나 평가하기 위해서는 아미노산 서열의 각각의 아미노산이 어떤 이차 구조에 속하는지를 나타낸 자료가 필요하다. 실험적으로 결정된 단백질 삼차원 구조에 이차 구조를 할당하기 위해서, 이차 구조의 분류로 가장 널리 사용되는 DSSP[13] 프로그램을 사용하였다. DSSP는 단백질 서열의 각 아미노산을 여덟 가지 (G: 3-helix, H: alpha helix, I: 5-helix, B: residue in isolated beta-bridge, E: extended strand, participates in beta ladder, T: hydrogen-bonded turn, S: bend, “.”: otherwise) 중의 하나로 분류한다. 또한, 이들 여덟 가지를 세 가지 (G, H, I -> H; B, E -> E; T, S, “.” ->C)로 통합하여 나선, 베타-병풍과 코일구조로 나눈다.

유사한 서열을 찾기 위한 대규모 단백질 자료는 질의 서열을 포함하고 있지 않은 121,713개의 단백질 체인으로 구성되었는데, 단백질 체인은 1개 이상 모여서 단백질 하나를 구성하는 단백질 구조의 일부분이다[3]. 질의 서열과 유사한 서열을 보다 광범위하게 탐색하기 위하여 갭을 허용하였고, blastpgp에서 세 번 반복하여 프로파일 구성할 때의 문턱치를 조절하였다.

Table. 1 Coverage, specificity, and accuracy of protein secondary structure prediction for *h*-values (%)

	Q3			Q8		
No Gap	79.97	96.18	76.92	79.24	93.29	73.92
Gap, $h = 10^{-30}$	88.54	95.22	84.31	88.54	91.71	81.20
Gap, $h = 10^{-20}$	88.54	95.22	84.31	88.54	91.71	81.20
Gap, $h = 10^{-10}$	88.62	95.21	84.38	88.62	91.68	81.25
Gap, $h = 10^{0}$	91.24	94.39	86.12	91.24	90.60	82.66

표 1은 이차 구조의 종류를 각각 3개와 8개를 분류한 Q3와 Q8의 이차 구조 예측 결과를 나타낸다. 질의 서열과 한 개 이상의 유사한 단백질 자료가 존재하여 이차 구조가 예측되는 비율(coverage), 이러한 예측이 수행된 범위 내에서의 정확도인 특이성(specificity), 모든 자료에 대해서 올바른 이차 구조 예측이 수행된 정도인 정확도(accuracy)를 차례로 표시하였다. 표 1의 두 번째 행에 나타난 No Gap은 SSpro[3]와 같은 조건인 질의 서열과 최소한 10개의 아미노산이 갭이 없이 일치하고, 10^{-9}이하의 e값을 갖고, 45%이상의 아미노산이 일치하고, 55%이상의 양의 치환값을 가지는 서열의 정보를 사용할 때의 결과이다. 표 1의 나머지는 탐색을 위한 프

로파일을 구성할 때의 옵션 *h*를 달리할 때의 결과를 나타내는데, 갭을 허용하고, 10^{-3}이하의 e값을 갖는 서열을 식 (1)의 $w=\exp(0.14)$으로 조합한 결과이다.

표 1에서 보듯이 갭을 허용하지 않는 No Gap은 전체 아미노산 서열의 79.97%만이 해당하는 이차 구조 정보를 단백질 자료에서 찾을 수 있으나, 갭을 허용하는 탐색은 최대 91.24%까지의 더 넓은 예측 범위를 가지고 있다. 더 많은 아미노산 서열에 대해서 예측이 가능하므로 전체적으로 정확도가 향상된다. 하지만 질의 서열과 정확히 일치하지 않은 부분의 부분서열도 사용하므로 특이성은 약간 하락하였다. 또한, 옵션 *h*가 클수록 프로파일의 구성에 포함되는 서열이 많아지므로, 질의 서열과 유사한 서열이 더 많이 탐색되어, 더 많은 아미노산의 위치에서 예측을 수행할 수 있는 범위와 정확도는 증가한다. 하지만, 예측된 것들 중에서 정확하게 예측된 비율인 특이성은 오히려 감소하는 상충관계(trade-off)가 나타난다.

Table. 2 Coverage, specificity, and accuracy of protein secondary structure prediction for cut-off *e*-values (%)

	Q3			Q8		
$e=10^{-4}$	88.05	95.31	83.92	88.05	91.81	80.84
$e=10^{-3}$	88.54	95.22	84.31	88.54	91.71	81.20
$e=10^{-2}$	89.06	95.14	84.73	89.06	91.57	81.56
$e=10^{-1}$	89.81	94.95	85.27	89.81	91.31	82.01
$e=10^{-0}$	90.59	94.64	85.73	90.59	90.92	82.36

표 2에서는 $h=10^{-20}$이고 식 (1)의 $w=\exp(0.14)$일 때, 여러 문턱치 *e*값을 사용한 결과이다. 문턱치가 커짐에 따라서 이차 구조 예측의 범위와 정확도는 증가하지만, 예측된 것들 중에서 정확하게 예측된 비율인 특이성은 오히려 감소한다. 이러한 상충관계는 표 1에서 *h*가 증가함에 따라 나타나는 경향과 일치한다.

표 1과 표 2에서는 질의 서열의 많은 부분을 예측할수록 이차 구조 예측의 정확도가 더불어 높아졌다. 이는 유사 서열이 존재하지 않아서 예측을 수행할 수 없는 범위가 10%정도로 크게 존재하기 때문이다. 하지만, 모든 질의 서열의 위치를 예측할 수 있는 방법들과 조합하면 이러한 예측 범위와 정확도의 동일한 경향성은 유지되지 않을 것이다. 이를 확인하기 위하여 본 논문의 제안 방법을 먼저 적용하고, 예측되지 않는 나머지 범위에 대하여 신경망을 사용하는 SSpro[3]의 예측결과를 적용하였다. SSpro는 질의 서열의 모든 위치에서 예측을 하는데, 본 논문의 실험 자료에 적용한 결과 Q3에서는 79.21%, Q8에서는 66.77%의 정확도를 보였다. 제안한 방법의 성능을 식 (2)의 넓은 파라미터 범위에서 조사하였다.

$$\begin{aligned} h &= 10^{-30}, 10^{-28}, 10^{-26}, \ldots, 10^{0}, \\ e &= 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, \\ w &= \exp(0.10), \exp(0.11), \exp(0.12), \ldots, \exp(0.4) \end{aligned} \quad (2)$$

제안한 방법의 예측 범위와 특이성의 상충관계로 인하여 표 1과 표 2와는 다른 *h*와 *e*에서 최고의 정확도를 보였다. Q3은 $h=10^{-18}$, $e=10^{-3}$, $w=\exp(0.14)$에서 93.28%의 정확도를 보였고, Q8은 $h=10^{-28}$, $e=10^{-2}$, $w=\exp(0.14)$에서 88.79%의 정확도를 보였다. 파라미터들이 $h=10^{-30}, 10^{-20}, 10^{-10}$, $e=10^{-3}, 10^{-2}$, $w=\exp(0.10), \exp(0.11), \ldots, \exp(0.18)$ 의 넓은 범위에서 평균을 구하니, Q3은 93.26%, Q8은 88.78%의 정확도로 성능이 안정적이었다. 신경망을 사용하여 모든 위치에서 이차 구조의 예측이 가능한 방법을 동시에 사용할 경우에는 *h*와 *e* 값이 작은 것을 사용하여 특이성이 큰 것을 사용하는 것이 더욱 효과적인 결과를 보였다.

표 3은 본 실험과 동일한 실험 자료를 사용하는 다른 방법들과의 비교이다. 본 논문의 방법이 가장 높은 정확도를 보임으로서, 갭을 허용하여 유사한 서열을 광범위하게 탐색한 후에 이를 상동관계의 정도에 따라 가중하여 조합하는 방법이 효과적임을 보여준다.

Table. 3 Comparison of accuracy of protein secondary structure prediction methods (%)

	Q3	Q8
Proposed	93.28	88.79
SSpro[3]	92.91	87.92
DISTILL Porter 4.0[4]	-	82.56
PSIPRED 3.3[2]	-	80.59

Ⅳ. 결 론

본 논문은 대규모의 단백질 구조 자료에서 정보를 효과적으로 추출하여 단백질 이차 구조를 예측하였다. 갭이 있는 서열 탐색과 질의 단백질과 상동 관계에 있는

단백질 자료내의 단백질들을 기반으로 질의 서열을 재구성하고 이를 이용하여 보다 먼 상동 관계의 서열들을 탐색하였다. 이러한 탐색과정에서 나타나는 서열의 유사성을 기준으로 최종 예측 결과에 기여하는 정도를 조절하였다.

기존의 방법에 비하여, 제안한 방법은 상동 관계가 있는 아미노산 서열을 더 광범위하게 사용하므로, 질의 서열의 더 많은 부분이 상동 관계를 이용하여 예측이 가능하여 정확도가 향상된다. 하지만 질의 서열과 정확히 일치하지 않은 부분의 부분서열도 사용하므로 특이성은 약간 하락하는 상충관계가 나타난다. 따라서 이러한 상충관계를 관계를 개선하기 위해서 서열의 상동 관계의 정도에 종속적인 예측방법의 구성이 필요하다.

최근 들어 구축되는 단백질 데이터베이스 규모가 매우 증대되고 있으므로, 임의의 질의 서열과 상동 관계의 데이터베이스내의 단백질 존재할 가능성이 점점 커지고 있다. 따라서 상동 관계를 보다 효율적으로 이용하는 방법의 개발이 필요하다.

REFERENCES

[1] H. Lodish, A. Berk, C.A. Kaiser, et al., *Molecular Cell Biology*, 6th Ed. New York, NY: W. H. Freeman and Company, 2007.

[2] H. W. Buchan, *et al.*, "Scalable web services for the PSIPRED protein analysis workbench," *Nucleic Acids Res.*, vol. 41, W72-W76, Jul. 2013.

[3] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, vol. 30, no. 18, pp, 2592-2597, Sep. 2014.

[4] C. Mirabello and G. Pollastri, "Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility," *Bioinformatics*, vol. 29, no. 16, pp. 2056-2058, Aug. 2013.

[5] R. Yan, *et al*, "A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction," *Scientific Reports 3*, Article number: 2619, Sep. 2013.

[6] J. Zhou and O. Troyanskaya, "Deep supervised convolutional generative stochastic network for protein secondary structure prediction," in *JMLR Proceedings*, 32, pp. 745- 753, Beijing, China, 2014.

[7] M. Spencer, J. Eickholt and J. Cheng, "A deep learning network approach to ab initio protein secondary structure prediction," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 12, no. 1, pp. 103-112, Jan.-Feb. 2015.

[8] R. Heffernan, *et al*, "Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning," *Scientific Reports 5*, Article number: 11476, June 2015.

[9] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Scientific Reports 6*, Article number: 18962, Jan. 2016.

[10] S. F. Altschul, *et al*, "Gapped blast and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389-3402, Sep. 1997.

[11] B. E. Suzek, *et al*, "Uniref: comprehensive and non-reduncant uniprot reference clusters," *Bioinformatics*, vol. 23, no. 10, pp. 1282-1288, May 2007.

[12] H. M. Berman, *et al*, "The protein data bank," *Nucleic Acids Res.* vol. 28, no. 1, pp. 235-242, Jan. 2000.

[13] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrohen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577-2637, Dec. 1983.

지상문(Sang-Mun Chi)

1991년 서울대학교 수학교육학과 졸업(이학사)
1993년 한국과학기술원 수학과 졸업(이학사)
1998년 한국과학기술원 전산학과 졸업(공학박사)
1993년 ~ 2000년 삼성전자 무선사업부 선임연구원
2001년 ~ 현재 경성대학교 컴퓨터공학과 교수
※관심분야 : 생물정보학, 기계학습