JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 22, No. 1 : 26~32 Jan. 2018

단백질 이차 구조 예측을 위한 단백질 프로파일의 성능 비교

지상문*

A Performance Comparison of Protein Profiles for the Prediction of Protein Secondary Structures

Sang-Mun Chi*

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

요 약

단백질의 이차구조는 단백질의 진화, 구조, 기능을 연구하는데 중요한 정보이다. 단백질 서열 정보만을 이용하여 단백질의 이차 구조를 예측하는 분야에 심층 학습 방법들이 최근 들어 활발히 적용되고 있다. 이러한 방법에서 널리 사용되는 입력은 단백질 서열을 변환하여 만들어진 단백질 프로파일이다. 본 논문에서는 효과적인 단백질 프로파일을 얻기 위하여 단백질 서열 탐색 방법으로 PSI-BLAST와 더불어서 HHblits를 사용하였다. 단백질 프로파일의 구성에 사용되는 상동 단백질 서열을 결정하기 위한 유사도 문턱치와 상동 단백질 서열 정보를 반복적으로 사용하는 회수를 조절하였다. 합성곱 신경망과 순환 신경망을 사용하여 단백질 이차구조를 예측하였는데, 진화적 정보를 한번만 추가하여 만들어진 단백질 프로파일이 효과적이었다.

ABSTRACT

The protein secondary structures are important information for studying the evolution, structure and function of proteins. Recently, deep learning methods have been actively applied to predict the secondary structure of proteins using only protein sequence information. In these methods, widely used input features are protein profiles transformed from protein sequences. In this paper, to obtain an effective protein profiles, protein profiles were constructed using protein sequence search methods such as PSI-BLAST and HHblits. We adjust the similarity threshold for determining the homologous protein sequence used in constructing the protein profile and the number of iterations of the profile construction using the homologous sequence information. We used the protein profiles as inputs to convolutional neural networks and recurrent neural networks to predict the secondary structures. The protein profile that was created by adding evolutionary information only once was effective.

키워드 : 단백질 이차 구조, 단백질 프로파일, 단백질 서열 탐색, PSI-BLAST, HHblits.

Key word : Protein secondary structure, Protein profile, Protein sequence search, PSI-BLAST, HHblits.

Received 24 August 2017, **Revised** 01 September 2017, **Accepted** 28 September 2017

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

Open Access http://doi.org/10.6109/jkiice.2018.22.1.26 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

단백질의 삼차원 구조와 기능을 결정하는 가장 큰 요소는 단백질을 구성하는 아미노산 서열로 알려져 있다[1]. 따라서 단백질 서열정보만을 사용하여 단백질의 구조와 기능을 예측하는 연구가 활발히 진행되고 있는데, 단백질의 이차 구조 예측은 이러한 연구들의 바탕을 이루는 중요한 요소 기술이다. 단백질의 일차 구조는 단백질을 구성하는 아미노산 서열을 의미하고, 이차 구조는 아미노산들이 국부적으로 나타나는 나선이나 병풍형태 등의 규칙적 구조를 의미한다[2]. 최근에는 심층 신경망(deep neural networks)을 이용하여 단백질 이차 구조 예측에 적용하려는 시도가 활발하다[3-7]. 본 논문에서는 심층 신경망 구조 중에서 성능이 우수한 합성곱 신경망(Convolutional Neural Network, CNN)과 순환 신경망(Recurrent Neural Network: RNN)을 사용하여 단백질 이차 구조를 예측한다[8,9]. 합성곱 신경망은 입력 자료의 국부적 관계를 효과적으로 모델링할 수 있으므로, 이미지, 영상, 음성, 오디오 처리에 매우 좋은 성능을 보여 주고 있으며[8-10], 순환 신경망은 순차적 자료의 처리에 적합하며, 이전 시간의 정보들을 효과적으로 학습하기 위해 개발된 긴 단기 기억(Long Short-Term Memory: LSTM)[11]과 게이트 순환 신경망(Gated Recurrent Unit: GRU) [12]이 대표적이다.

최근에는 대부분의 단백질 이차 구조를 예측하는 방법들이 단백질 프로파일을 입력으로 사용한다. 단백질 프로파일은 단백질 데이터베이스에서 유사한 서열들의 정보를 이용하여 변환하려는 단백질 서열의 각 위치에서 20개 아미노산의 치환 빈도로 구성된다. 이러한 서열의 위치별로 치환빈도를 표시하는 행렬이 단백질 구조를 보다 잘 나타내므로 단백질의 특징을 예측하고 비교하는 분야에 널리 사용된다[13,14]. 본 논문에서는 단백질 프로파일을 구성하기 위해서 PSI-BLAST[13]와 HHblits [14]를 사용하였다. 단백질 프로파일을 구성하기 위하여 단백질 데이터베이스로부터 추가되는 서열을 결정하는 유사도 문턱치와 상동 단백질의 정보를 이용하여 반복적으로 프로파일을 구성하는 횟수에 따른 성능을 조사하였다. HHblits을 사용하여 단백질 프로파일로 변환하기 위하여 HHblits이 출력 결과를 PSI-BLAST와 같은 형태로 변환한다. 단백질 프로파일들을 합성곱 신경망과 순환 신경망의 입력으로 사용하여 최적의 성능을 보이는 단백질 프로파일을 찾는다.

2장에서 본 논문의 실험에 사용되는 심층신경망의 구조와 파라미터를 살펴보고, 3장에서 단백질 프로파일을 구성하는 방법을 논의하고, 4장에서 심층신경망과 단백질 프로파일을 사용하여 단백질의 이차 구조를 예측하는 실험을 하고, 5장에서 결론을 맺는다.

Ⅱ. 심층신경망의 구조

심층 신경망은 여러 계층으로 구성되어져 있어서 복잡한 자료를 모델링하기에 적합한 구조이다. 즉, *l*-번째의 계층의 *i*-번 유닛 o_i^l은 (*l-1*)-번째 계층의 유닛들의 결과를 식 (1)과 같이 변환하여 얻는다. 처음 계층의 유닛들 o_i^0는 입력 자료이다.

$$o_i^l = f(\sum_j o_j^{l-1} w_{j,i}^l + w_{0,i}^l) \quad (l = 1,2,...,L-1) \quad (1)$$

여기서, $w_{j,i}^l$은 (*l-1*)-번째 계층의 *j*-번 유닛과 *l*-번째 계층의 *i*-번 유닛을 연결하는 가중치, $w_{0,i}^l$는 *l*-번째 계층의 *i*-번에 더해지는 편향이다. 함수 f는 일반적으로 $\mathrm{relu}(x) = \max(0, x)$, $\sigma(x) = 1/(1+\exp(-x))$나 $\tanh(x) = (1-\exp(-2x))/(1+\exp(-2x))$이다. 분류를 수행하는 문제에서는 마지막 계층의 변환함수는 f대신에, 이전 계층의 출력에 식 (2)의 소프트맥스(softmax) 함수를 적용하여 각 부류의 사후 확률 y_i을 얻는다.

$$y_i = \exp(o_i^L) / \sum_j \exp(o_j^L) \quad (2)$$

합성곱 신경망은 생명체의 수용체가 받은 자극을 인접한 부위에만 영향을 준다는 실험적인 사실을 이용하여 인접한 자료간의 상관관계를 모델링하기에 적합하도록 만들어졌다. 이러한 국부적 연결을 나타내는 파라미터를 전체 자료의 각 부분 영역에 동일하게 공유하여 위치의 변화에 무관한 불변성을 얻는다. 본 논문의 CNN은 (1) 입력으로 이차원의 단백질 프로파일 $o_{i,j}$를 사용한다. 첨자 i는 서열의 위치이고 j는 20차원으로 각 아미노산의 치환 빈도이다. (2) $h_{i,j}^{m,k}$를 m 계층의 k번째 특징지도라면 $m-1$ 계층의 여러 특징 지도를 이용하여 얻는다.

$$h_{i,j}^{m,k} = f(\sum_{n=1}^{F_{m-1}} \sum_{u=-T/2}^{T/2} \sum_{v=-L/2}^{L/2} h_{u,v}^{m-1,n} w_{i-u,j-v}^{k,n} + w_m^k) \quad (3)$$

단, F_{m-1}는 $m-1$ 계층의 특징 지도들의 개수, T와 L은 계산에 사용되는 국부영역의 크기이며, w들은 학습하여야 할 파라미터이다. 입력 단백질 프로파일 $o_{i,j}$을 $h_{i,j}^{0,0}$로 하고 이때의 특징지도는 1개이다. (3) 통합 계층은 아래 계층의 특징 지도의 평균값 또는 최대값을 구하여 특징을 요약한다. 단백질 이차 구조의 예측 실험에는 효과가 없어서 본 논문에서는 적용하지 않았다. (4) 마지막 계층의 출력을 일차원으로 변환하여 식 (1)의 신경망의 입력으로 사용하고, 식(2)로 최종 부류별 확률을 구한다. 순환 신경망은 순차적 자료의 처리에 적합하도록 구성되어 있다. 입력 서열 $x=(x_1, x_2, ..., x_T)$로부터 은닉 벡터열 $h=(h_1, h_2, ..., h_T)$와 출력 $y=(y_1, y_2, ..., y_T)$를 $t=1$부터 식 (4)-(5)를 차례로 수행하여 얻는다.

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (4)$$

$$y_t = W_{hy}h_t + b_y \quad (5)$$

단, W들과 b는 학습을 통하여 얻을 가중치 행렬과 편향벡터이고, H는 $\sigma()$나 $\tanh()$가 사용된다.

RNN은 학습 과정 중에 전파되는 편미분 값이 매우 작아지거나, 급격히 커지는 현상이 서열의 길이가 긴 경우에 발생하는데, 이러한 경우에는 학습이 어렵다. 이를 해결하고자 LSTM[11]은 여러 개의 게이트를 이용하여 이전 벡터값을 얼마만큼 사용할 지를 결정한다.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (6)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (7)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (8)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \quad (9)$$

$$h_t = o_t \tanh(c_t) \quad (10)$$

단, i, f, o, c를 각각 입력 게이트, 까먹음 게이트, 출력 게이트, 셀 활성화 벡터라 부른다. 셀에서 게이트를 연결하는 가중치 행렬들(예, W_{cf})은 대각행렬이다. 또 다른 방법인 GRU[12]의 구조는 다음과 같다.

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1}) \quad (11)$$

$$u_t = \sigma(W_{xu}x_t + W_{hu}h_{t-1}) \quad (12)$$

$$c_t = \sigma(Wx_t + r_t \diamond (W_{hc}h_{t-1})) \quad (13)$$

$$h_t = (1-u_t) \diamond h_{t-1} + u_t \diamond c_t \quad (14)$$

단, $\diamond$는 원소별 곱셈이고, 리셋 게이트 r_t는 새로운 입력을 이전 메모리와 어떻게 합칠지를 정해주고 업데이트 게이트 u_t는 이전 메모리를 얼마만큼 기억할지를 결정한다. GRU는 LSTM보다 파라미터 개수가 적어, 학습 시간이 짧고 적은 자료로 학습이 가능하나, 복잡한 자료를 모델링하는 능력은 상대적으로 작다.

본 논문에서는 보다 복잡한 상관관계를 모델링하기 위해 BRNN(Bidirectional Recurrent Neural Network)과 DRNN(Deep Recurrent Neural Network)을 사용하였다[15,16]. BRNN은 식 (4)의 h_t와 동일한 구조이지만 $t=T, T-1, ..., 1$의 역방향으로 은닉 벡터 g_t를 구하고, h_t와 g_t를 이용하여 출력 벡터 y_t를 구한다.

$$g_t = H(W_{xg}x_t + W_{gg}g_{t+1} + b_g) \quad (15)$$

$$y_t = W_{hy}h_t + W_{gy}g_t + b_y \quad (16)$$

DRNN은 RNN의 은닉 계층 위에 새로운 은닉 계층을 쌓는 방법이다. 즉, N개의 계층으로 이루어진 DRNN은 RNN의 식 (4)를 다음과 같이 변경한다.

$$h_t^n = H(W_{h^{n-1}h^n}h_t^{n-1} + W_{h^n h^n}h_{t-1}^n + b_h^n) \quad (15)$$

$$y_t = W_{h^N y}h_t^N + b_y \quad (16)$$

단, 각각의 t에서 $h_t^0 = x_t$이고, $n=1, 2, .., N$ 순서로 계산한다. DBRNN(Deep Bidirectional Recurrent Neural Network:)은 BRNN을 여러 개 쌓은 DRNN이다.

신경망의 파라미터를 학습하기 위해, 최소화하는 목적함수로서 음의 로그우도를 사용하였다.

$$l(T, Y) = -\sum_{i=1}^{n}\sum_{c=1}^{C} t_{ic}\log(y_{ic}) \quad (17)$$

단, n은 자료의 수, C는 클래스의 수, t_{ic}는 i번째 자료의 클래스가 c일 경우에만 1이고 나머지는 0이고, y_{ic}는 i번째 자료의 클래스가 c일 출력확률이다. 신경망의 학습에는 각 부류의 사후 확률 $y_{ic}(c=1,...,C)$와 목적함수 (식 17)를 사용하여 다음 편미분 값을 구한다.

$$\partial l / \partial o_{ic}^L = y_{ic} - t_{ic} \quad (18)$$

식 (18)부터 역전파 학습[8-12]으로 각 계층의 파라미터의 편미분을 계산하여 파라미터의 갱신에 이용한다.

Ⅲ. 단백질 프로파일 구성

단백질 프로파일은 단백질의 구조와 기능을 예측하는 많은 방법들의 중요한 정보로 사용되며, 본 논문에서도 심층 신경망의 입력 자료로 사용하였다. 단백질 프로파일은 단백질 데이터베이스에서 유사한 서열들과의 정합(alignment)으로부터 얻은 여러 개의 유사 서열의 정보를 집약하여 나타내는 방법이다[13,14]. 유사성이 높은 아미노산 서열들은 단백질 구조가 유사하므로, 유사한 단백질 서열들의 특징을 넓게 포함한 단백질 프로파일이 하나의 단백질 서열보다 더 많은 정보를 표현할 수 있는 장점이 있다.

단백질 프로파일을 구성하기 위해 널리 사용하는 방법은 PSI-BLAST[13]이다. 이 방법은 변환하려는 단백질 서열과 유사한 서열들을 단백질 데이터베이스에서 탐색하여 PSSM (position specific scoring matrix)을 구성한다. PSSM은 단백질 서열의 길이가 N이라면 이들 각각의 위치에서 20개의 아미노산이 관측될 확률을 유사한 서열들과의 정합을 통하여 구한 것으로 $N\times 20$의 차원의 행렬이다. PSI-BLAST는 단백질 서열대신에 PSSM을 입력으로 사용하여 반복적으로 PSSM을 구함으로서 초기의 입력 단백질 서열만으로는 찾을 수 없었던 상동 단백질 서열들의 정보를 포함할 수 있다. 하지만 이러한 반복의 단점으로는 질의 단백질의 특성과 유사하지 않는 특징이 계속 추가되어 누적될 수 있다는 위험성이 존재한다. 단백질 데이터베이스는 변환하려는 단백질 서열과 유사한 서열들을 찾는데 이용된다. 본 논문에서는 PSI-BLAST를 사용한 탐색에는 대용량 단백질 서열 자료인 UniRef50[17]을 이용하였다. 주어진 유사도보다 큰 유사성을 갖는 서열만을 사용하여 PSSM을 구성하는데, 이러한 문턱치 E값에 따라 최종적으로 만들어지는 단백질 프로파일의 특성이 달라진다. 즉, E값은 우연에 의하여 데이터베이스내의 서열들끼리 정합되는 개수이므로, 작은 E값을 문턱치로 사용할 경우에는 매우 유사한 서열들만을 사용하여 단백질 프로파일을 만들게 된다. 본 논문에서는 최적의 단백질 프로파일을 조사하기 위하여 psi-blast 프로그램의 문턱치 E값 (기본값 10)를 결정하는 [-evalue] 옵션을 사용하여 다음의 4가지 경우에 대하여 각각 단백질 프로파일을 구하였다.

$$-evalue = 0.1, 1, 10, 100 \tag{19}$$

또한, [-num_iterations] 옵션을 6으로 설정하여 PSI-BLAST가 반복적으로 PSSM을 5개까지 구하도록 하였다. 최종적으로 만들어지는 단백질 프로파일은 4개의 E값 각각에 대하여 5개의 반복으로 만들어진 20개이다. 신경망의 입력으로는 이렇게 얻어진 PSSM에 0.1을 곱하여 입력 값의 크기를 조절하여 사용하였다.

단백질 프로파일을 구성하는 또 다른 방법인 HHblits[14]은 HMM(hidden Markov model)으로 프로파일을 만드는데, PSSM처럼 단백질 서열의 각 위치에서 아미노산이 관측되는 확률을 가지고 있고, 추가적인 파라미터로 각 위치에 서열이 삽입되거나 삭제되는 확률도 가지고 있다. HHblits에서도 단백질 서열을 단백질 프로파일을 변환하기 위해서 유사한 서열들을 탐색하여야 하는데, 이를 위하여 단백질 데이터베이스로 UniProt[18]을 사용하였다. HHblits도 PSI-BLAST처럼 입력 서열 대신에 다중 서열이 정합된 결과를 입력으로 사용하여 반복적으로 단백질 프로파일을 만들 수 있는데 본 논문에서는 5번의 반복을 수행하고 매 반복마다의 프로파일을 구하였다. HHblits도 PSI-BLAST의 문턱치 E값처럼 각 반복마다 추가되는 유사서열을 [-e] (기본 값 0.1) 옵션으로 조절할 수 있다. 본 논문은 다음의 4개의 값을 사용하였다.

$$-e = 0.001, 0.01 0.1, 1 \tag{20}$$

단백질 프로파일을 구성하기 위해서 HMM의 각 상태에서의 여러 아미노산의 관측확률만을 사용하였는데, HHblits은 관측확률은 다음의 형태로 출력한다.

$$-1000 \times \log_2(frequency) \tag{21}$$

본 논문에서는 식 (21)에서 20개 아미노산의 평균적인 빈도인 1/20에서 단백질 프로파일의값이 0이 되도록 식 (22)로 변환하였고, -3.5보다 작은 값들은 -3.5로 만들어서 너무 적은 빈도간의 차이는 무시하였다.

$$\log_2(frequency) + \log_2 20 \tag{22}$$

Ⅳ. 단백질 이차구조 예측 실험

4.1. 실험 환경

실험을 위한 자료를 수집하기 위하여 CullPDB server[19]에서 30%이하의 서열 동일성을 갖고 단백질 구조의 해상도가 2.5 옹스트롬 보다 양호한 단백질 자

료를 대상으로, 논문 [5,7]과 같이 길이가 50이상이고 700이하인 단백질 체인을 선택하였고, 체인의 개수는 9091개이다. 단백질 서열로부터 예측한 단백질 이차구조의 정확도를 판단하기 위해서는, 실제 단백질 삼차원 구조로부터 얻은 이차구조가 필요하다. 단백질 구조자료를 DSSP[20] 프로그램을 사용하여 8개의 이차구조 (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 분류하였고, 이를 세 가지 (G, H, I -> H; B, E -> E; T, S, "." ->C)의 대분류인 나선(H), 베타-병풍(E)과 코일(C)로 변환하여 본 논문의 이차 구조를 정의하였다. 심층 신경망을 구성하고 있는 여러 계층은 복잡한 자료의 관계를 모델링할 수 있는 반면에 학습 자료에 나타난 특성에 지나치게 적응할 수 있다. 학습 자료에만 과적응된 신경망은 실제 평가 자료의 예측에는 효과적이지 않다. 이를 방지하기 위하여 조기중단(early stopping)을 사용하였다. 학습을 통하여 파리미터를 갱신하고, 학습 자료와 별개의 검증 자료를 예측하고 검증자료의 정확도가 현재까지의 갱신 회수의 2배의 회수의 학습과정에서 정확도가 향상되지 않으면 학습을 중단한다. 본 논문에서는 실험 자료를 균등하게 5개로 나누고, 1개는 평가 자료로, 1개는 검증 자료로, 나머지 3개는 학습 자료로 사용하였다.

심층신경망으로 이차 구조 예측을 수행하기 위하여 Theano[21]와 Lasagne[22] 패키지를 사용하여 CNN과 RNN을 구현하였다. 신경망의 파라미터를 갱신하는 학습 방법으로는 Adagrad[23]를 사용하였다.

4.2. 최적 단백질 프로파일 선택

단백질 이차구조 예측에 적합한 단백질 프로파일을 찾기 위하여 여러 단백질 프로파일을 심층 신경망의 입력 자료로 사용하여 예측 정확도를 비교하였다.

표 1과 표 2는 CNN을 사용하여 단백질 이차구조를 예측한 정확도이다. CNN의 구조는 식 (3)에서 입력 자료 $h_{i,j}^{0,0}$는 예측하려는 아미노산서열의 위치를 중심으로 하여 주변의 39개의 아미노산 부분 서열에 해당하는 단백질 프로파일이다. 1 번째 계층의 구조는 $F_1 = 160$, $T = 5$, $L = 20$이고, 2번째 계층은 $F_2 = 160$, $T = 5$, $L = 1$이다. 2번째 계층의 모든 출력 4960개를 입력으로 식 (1)의 계층과 식 (2)의 계층을 차례로 통과하여 최종 분류확률을 예측하였다.

학습과정에서 학습률을 0.001, 0.01, 0.1을 사용하여 검증 자료에 대해 최고의 예측정확도를 주는 학습률을 찾고, 이렇게 얻은 학습률의 2배와 1/2배의 학습률을 적용하여 다시 최고의 정확도를 보이는 학습률을 찾았다. 표 1은 학습률 0.005, 표 2는 학습률 0.01을 사용하였다. 표 1은 PSI-BLAST를 여러 E값과 반복을 수행하여 얻은 단백질 프로파일을 사용한 결과이다.

Table. 1 Accuracy of protein secondary structure prediction using PSI-BLAST and CNN (%)

Iteration \ E value	0.1	1	10	100
1	83.66	83.83	83.71	83.87
2	83.66	83.79	83.66	83.67
3	82.86	82.95	82.89	82.90
4	82.26	82.15	82.22	82.15
5	81.63	81.56	81.68	81.55

표 1에서 보듯이 단백질 프로파일을 구할 때에 사용한 반복의 횟수가 증가함에 따라 성능이 저하되었다. 이는 단백질의 구조를 예측하는 문제에서는 반복적으로 상동 단백질을 사용하여 단백질 프로파일을 구성할 경우에는 질의 단백질의 구조와 다른 구조를 갖는 단백질들 포함하여 질의 단백질의 특성과 다르게 표류(drift)하는 경향이 있기 때문이라 판단된다. 상대적으로 E값에 따른 차이는 크지 않았다. 아미노산 부분서열의 길이가 35와 43인 경우에도 최적 학습률은 0.005였고, 전체적인 정확도의 경향은 부분 서열의 길이가 39인 표 1과 같았다. 표 2는 HHblits을 여러 E값과 반복을 수행하여 얻은 단백질 프로파일을 사용한 결과이다.

Table. 2 Accuracy of protein secondary structure prediction using HHblits and CNN (%)

Iteration \ E value	0.001	0.01	0.1	1
1	83.48	83.30	83.50	83.48
2	82.97	82.79	82.85	82.74
3	82.39	82.10	81.90	81.85
4	81.69	81.69	81.42	81.10
5	81.58	81.27	81.10	80.56

표 2의 경우에도 표 1과 마찬가지로 반복을 사용할수록 성능이 저하되었고, E값에 따른 차이는 크지 않았다.

아미노산 부분서열의 길이가 35와 43인 경우에도 최적 학습률은 0.01로서 표 2와 같았고, 전체적인 정확도의 경향은 부분 서열의 길이가 39인 표 2와 같았다.

표 3과 4는 신경망으로 CNN대신에 RNN을 사용하고 표 1, 2와 동일한 단백질 프로파일로 실험한 결과이다. RNN의 구조는 식 (11)-(14)의 GRU에서 입력벡터 x_t는 예측하려는 위치와 이 위치의 앞과 뒤에 해당하는 단백질 프로파일로서 60차원이며, 은닉벡터 h_t의 차원은 120이다. 또한, 식 (15)의 역방향에서의 은닉벡터 g_t도 사용 하며 차원은 120이고, 식 (16)과 같이 연결하여 출력벡터 y_t를 만들었다. 이를 다시 입력으로 사용하여 전방향과 역방향의 GRU를 만들고 식 (16)으로 연결하여 출력벡터를 만든다. 이를 식 (6)-(10)의 LSTM을 이용하여 각 위치의 단백질 이차구조를 예측하였다. 최적 학습률은 표1과 2에서와 같은 방법으로 얻었고, 학습률 0.01을 사용하였다.

Table. 3 Accuracy of protein secondary structure prediction using PSI-BLAST and RNN (%)

Iteration \ E value	0.1	1	10	100
1	82.80	82.82	82.89	82.87
2	82.85	82.82	82.90	82.77
3	82.11	82.10	82.12	82.14
4	81.43	81.53	81.30	81.37
5	80.92	80.91	80.89	80.91

Table. 4 Accuracy of protein secondary structure prediction using HHblits and RNN (%)

Iteration \ E value	0.001	0.01	0.1	1
1	81.74	81.83	81.66	81.80
2	81.43	81.42	81.35	81.21
3	81.33	80.83	80.68	80.61
4	80.83	80.40	80.07	79.96
5	80.21	80.05	79.98	79.35

표 3과 4에서 보듯이 RNN을 사용할 경우에도 CNN을 사용한 표 1과 표 2의 결과와 같이 단백질 프로파일을 구할 때에 반복을 사용할수록 성능이 저하되었고, E 값에 따른 차이는 크지 않았다. 하지만 예측 정확도가 CNN에 비하여 RNN을 사용할 경우에 하락되었다. 즉, 이차 구조의 예측에는 인접한 정보를 차례로 이용하는 것보다, 넓은 영역의 단백질 프로파일의 정보를 활용하는 것이 보다 효과적이었다. 은닉벡터의 차원을 90과 150일 경우의 예측정확도는 표 1과 표 2의 경향과 동일하였다.

Ⅴ. 결 론

본 논문에서는 단백질 이차 구조의 예측을 위하여 효과적인 단백질 프로파일을 구성하였다. 단백질 프로파일을 구성하기 위하여 질의 단백질과 상동 관계에 있는 단백질 자료내의 단백질들을 탐색하는 방법으로 기존에 널리 사용되는 PSI-BLAST이외에 추가적으로 HHblits를 사용하였다. 여러 가지 단백질 프로파일을 최근 들어 성능이 급격히 향상된 심층 심경망인 합성곱 신경망과 순환 신경망의 입력으로 사용하여 예측 정확도를 조사하였다. 단백질 프로파일을 구성하기 위하여 추가되는 상동 단백질의 문턱치는 예측정확도에 미치는 영향이 크지 않았다. 하지만, 단백질 이차 구조와 같은 단백질 구조의 예측에 있어서는 상동단백질들을 여러 번 반복적으로 추가하는 만들어지는 단백질 프로파일은 예측 정확도가 감소하는 경향을 보였다.

향후에는 단백질 이차 구조의 예측에 최적화되도록 다양한 신경망 구조와 학습 알고리즘을 적용하고, 신경망들의 앙상블을 사용하면 예측 정확도를 더욱 향상시킬 수 있을 것이다.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A3B03935290).

REFERENCES

[1] D. Baker and A. Sali., “Protein structure prediction and structural genomics,” *Science*, vol. 294, pp. 93-96, Oct. 2001.

[2] H. Lodish, *et al.*, Molecular Cell Biology, sixth Ed., W.H. Freeman and Company, New York, 2007.

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

[4] 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, pp, 2592-2597, Sep. 2014.

[5] J. Zhou, and O. Troyanskaya, "Deep supervised convolutional generative stochastic network for protein secondary structure prediction," *Journal of Machine Learning Research W&CP*, vol. 32, pp. 745-753, Jun. 2014.

[6] M. Spencer, J. Eickholt, and J. Cheng, "A deep learning network approach to ab initio protein secondary structure prediction," *IEEE/ACM Transactions on Computational Biology Bioinformatics*, 12, pp. 103-112, Jan/Feb. 2015.

[7] S. Wang, *et al.*, "Protein secondary structure prediction using deep convolutional neural fields," *Scientific Reports 6*, Article number: 18962, Jan. 2016.

[8] J. Schimidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85-117, Jan, 2015.

[9] Y. LeCunn, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436-444. May 2015.

[10] O. Abdel-Hamid, *et al.*, "Convolutional Neural Networks for Speech Recognition". *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 10. pp. 1533-1545, Jul. 2014.

[11] A. Graves, *et al.*, "Generating sequences with recurrent neural networks," *arXiv preprint* 1308.0850, Jun. 2014.

[12] C. Kyunghyun, *et al.*, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint* 1409.1259, Oct. 2014.

[13] S. F. Altschul, *et al.*, "Gapped blast and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389-3402, Sep. 1997.

[14] M. Remmert, A. Biegert, and J. Soding, "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, pp. 173-175, Dec. 2011.

[15] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Cananda, May 2013.

[16] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, Vancouver, Canada, May 2013.

[17] B. E. Suzek, *et al.*, "Uniref: comprehensive and non-reduncant uniprot reference clusters," *Bioinformatics*, vol. 23, pp. 1282-1288, May 2007.

[18] The UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 45, D158-D169, Jan. 2017.

[19] G. Wang and R.L. Dunbrack "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, pp. 1589-1591, Aug. 2003.

[20] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrohen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577-2637, Dec. 1983.

[21] Theano Development Team. "Theano: A Python framework for fast computation of mathematical expressions," arXiv e-prints, 1605.02688, May 2016.

[22] S.. Dieleman, et al., "Lasagne: First release," DOI:10.5281/zenodo.27878, http://dx.doi.org/10.5281/zenodo.27878, Aug. 2015.

[23] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, Jul. 2011.

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