HMM을 이용한 단백질 β -barrel 막횡단 부위 예측

*안창신O ,*유성준 ,**박현석
*세종대 컴퓨터공학부, **이화여대 컴퓨터학과
rebeli24@hotmail.com, sjyoo@sejong.ac.kr, neo@ewha.ac.kr

Predicting Transmembrane β -barrel membrane protein with HMM

*Chang Shin AhnO ,*Seong Joon Yoo, **Hyun Seok Park
*School of Computer Engineering, Sejong University, **Dept. of Computer Science & Engineering, Ewha Womans University

요 약

2000년대 초 인간 지놈 프로젝트의 완성으로 새로운 포스트-지놈 시대를 맞이하여, 유전자에 대한 해독보다는 인간의 모든 대사와 질병에 직접관여 하고 있는 단백질의 구조와 기능에 대해 많은 관심과 연구가 이루어지고 있다. 특히, 특정 단백질들은 암과 같은 불치병에 직접관여 하고 있으므로 이러한 단백질들의 기능과 구조에 대한 예측 성능의 향상은 새로운 신약 개발에 큰 도움이 될 것이다. 본 논문은 기계학습(Machine Learning)의 한 분야인 HMM(Hidden Markov Model)을 이용하여 β -barrel 형태로 막횡단하는 단백질의 특성과 기능으로부터 막횡단하는 부위가 존재하는지 여부를 예측하는 프로그램을 구현했다.

1. 서 론

본 논문에서는 단백질 중 β -barrel 막횡단 부위가 존재하는 서열 부위를 예측 하는 알고리즘을 소개한다. α -helix 막횡단 단백질과 달리 β -barrel 막횡단 단백질 구조를 밝힌 data가 적기 때문에 이것이 예측방법 개발 가능성을 제한해왔다. 정확한 barrel의 위상을 찾기 위해서는 막횡단 지역의 위치를 정확하게 결정하는 것과 존재하는 템플릿들을 기본으로 그 위에서 3차원 모델을 만드는 것이 필요하다. 그러나, 이 작업은 α -helix 막횡단 단백질의 예측보다 어려운 것으로 나타났다[1][2][3]. 최근 NN(Neural Network)나 HMM(Hidden Markov Model)을 이용 β -barrel 막횡단에 대해서도 연구와 개발이 조금씩 이루어지고 있다. 본 논문에서는 그 중 HMM을 이용하여 시스템을 구현하였다.

2. β -barrel 막횡단 단백질

2.1 β -barrel 막횡단 단백질의 특징

현재, 막횡단 단백질은 크게 2가지의 형태로 특징지어진다. 첫 번째 형태의 막횡단 단백질들을 구별하는 주된 특징은 세포의 인지질 이중막 횡단 구조가 α -helix 형태의 2차 구조를 갖는다는 것이다[4]. 반면에 두 번째 형태의 막횡단 단백질들은 인지질 이중막의 바깥쪽 막에서 2차 구조로 통과한다는 것과 단량체나 소중합체로 존재하는 antiparallel β -strands을 가지고 상호 작용한다는 차이점을 보인다[5][6]. 이들은 병풍(barrel) 모양의 구조를 가지며 인지질 쪽은 소수성, 병풍의 안쪽 통로는 친수성의 특징을 가지며 그로 인해 이온들과 작은 친수성(hydrophilic)의 molecule들의 수동적인 운반을 수행하거나, 전체 세포의 물질대사와 관련되는 기능들을 한다.

2.2 β -barrel 막횡단 단백질의 예측

β -barrel 막횡단 단백질 구조의 분석이 밝혀 진지 10년이 지났음에도 불구하고, β -barrel 막횡단 단백질로 알려진 단백질의 3차원 모델을 유추해 내는 것은 거의 불가능한 상태이다[5]. 왜냐하면 β -barrel 막횡단 단백질은 비록 같은 막횡단 지역에 있더라도 α -helix 막횡단 단백질과는 달리 소수성인 아미노산 외에도 친수성 아미노산을 포함하기 때문이다. 이런 이유로 α -helix 막횡단 단백질과는 달리 예측용 도구 개발이 미비했었다. 하지만 최근에 와서 NeuralNetwork, HMM을 이용한 연구와 개발이 조금씩 이루어지고 있다[7][8]. 본 논문에서는 HMM을 이용해서 단백질β -barrel의 막횡단 여부 예측 시스템을 구현했다.

3. HMM 모델을 이용한 막횡단 지역 예측

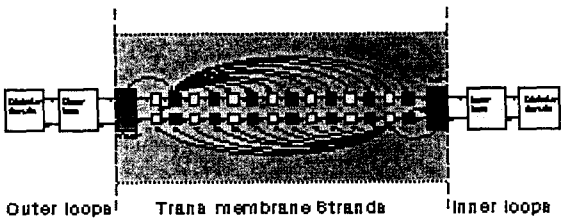
HMM은 바이오인포매틱스의 여러 분야에서 적용되고 있는 기계학습의 한 모델이다[9][10]. 이는 존재하는 실제 데이터(observed data)의 트레이닝을 통해서 그 데이터의 특성에 최적화된 모델을 만들고, 새로운 데이터에 대한 확률적 가능성을 제시 할 수 있는 특성을 가지고 있으며 유전자 및 단백질 서열 등에 적용하기에 적합한 특성을 지니고 있다. 본 논문에서는 β -barrel 막횡단 단백질들의 실제 데이터(20개의 알파벳으로 표현되는 아미노

산 서열)를 가지고 HMM 트레이닝을 통해 그에 대한 적절한 모델을 만들고 몇 가지 평가 방법을 통해 그 예측 결과에 대해 평가한다.

3.1 트레이닝 셋

2개의 논문에서 트레이닝 데이터로 사용되고 있는 것은 PDB(Protein Data Bank)로부터 선택되어진 11개의 데이터와, SWISSPROT으로부터 선택되어진 12개(PDB+1개추가)가 있다[7][8]. 이 데이터는 3차원 구조가 규명된 β -barrel 막횡단 단백질의 여러 종류 중 대표가 되는 단백질들이 선택되었다. 이 논문에서는 12개 중 6개의 훈련 예를 사용하였으며 6개의 위상에 대한 주석은 이전 연구자들의 것을 참고 하였다[7].

3.2 β -barrel HMM의 구조



[그림1] β -barrel HMM의 모델 구조

[그림1]은 현재 개발된 시스템의 모델을 나타내고 있다. 모델의 구조는 β -barrel 막횡단 단백질 데이터베이스로부터 얻어진 많은 β -barrel strands의 통계적인 특성들을 반영하고 있다[8]. 모델을 살펴보면 6가지의 state를 가지고 있으며, 각각의 state간에 반영되어진 트랜지션 확률을 가지고 순환을 하는 구조를 가지고 있다. 그 6가지 state는 barrel-strands를 위한 2개의 state(barrel-strands의 친수성과 소수성이 교대로 나타나는 특성), barrel-strands 양쪽의 inner와 outer loop에 접하는 부분의 barrel-strands cap 1개, outer loop에 대해 1개, inner loop에 대해 1개, inner와 outer loop의 중간에 존재하는 globular domain에 대한 state 1개 이다. 그리고 이들 state내 각각의 state들의 총 개수는 전체 54개의 state를 갖는다. 위의 그림에서 6개의 state에 대해 같은 색으로 표시되고 있으며, 또한 이는 같은 아미노산 분산 확률을 나타내고 있다.

3.3 모델의 생성 및 예측방법

주어진 트레이닝 셋으로부터 그 특성에 최적화된 모델을 만들어 내야 하는데 기본적으로 거의 모든 HMM 들에서 사용하는 모델 트레이닝 방법인 Expectation-Maximisation 알고리즘의 일반화된 형태를 기본적으로 사용한다[11]. 이 단계에선 HMM에서 가장 중요한 2가지 파라미터인 트랜지션(transition) 확률과 분산(emission) 확률의 업데이트가 이루어진다. 특히 본 논문에서는 labeling sequence 알고리즘을 이용하여 모델 업데이트 시에 파라미터로서 추가로 적용하여 계산하는 방법을 사용한다[12]. 트레이닝이 끝난 후 실제 데이터로부터 만들어진 최적화된 모델을 가지고 예측을 하는데, Viterbi algorithm을 사용한다[11].

4. 평가 방법 및 예측결과 분석

예측 시스템에 대한 평가는 통계적인 지표들을 가지고 시스템의 예측 결과에 대한 확률적인 수치들을 통해 이루어진다. 현재 β -barrel 막횡단 단백질의 평가에 사용되고 있는 지표들은 다음과 같다[10].

Q(2):전체 residues 의 개수 중 정확히 예측된 residues 의 개수

Q(β): barrel strand 부분에 대한 sensitivity(tp/(tp+fn))

Q(c): non-barrel strand 부분에 대한 sensitivity (tp/(tp+fn))

P(β): barrel strand 부분에 대한 specificity(tp/(tp+fp))

P(c) : non-barrel strand 부분에 대한 specificity(tp/(tp+fp))

C(β) : correlation coefficient = $((tp*tn-fp*fn)/(tp+fn)*(tp+fp)*(tn+fn)*(tn+fp)^{1/2})$

Sov(β) : barrel strand의 segment당 예측 정확도[13].

[표1] 예측 결과 평가표

β:barrel strand, c:non-barrel strand, Q: sensitivity, P: specificity

	Q2	Q(β)	Q(c)	P(β)	P(c)	C(β)	Sov(β)
Training	0.80	0.83	0.79	0.84	0.78	0.62	0.92
Testing	0.65	0.72	0.65	0.75	0.62	0.37	0.79

[표1]은 6개의 β -barrel 막횡단 단백질의 트레이닝 데이터를 가지고 [그림1]의 β -barrel HMM 모델 위에서 트레이닝 한 다음 예측 결과를 보여주고 있다. Training 필드는 6개의 트레이닝 데이터로 트레이닝 후 그 6개의 데이터에 대한 예측을 했을 때의 지표를 보여주며, Testing 필드는 cross-validation 기법을 적용했을 때의 지표를 보여준다. 즉, 5개의 아미노산 서열로 트레이닝한 모델을 가지고 나머지 하나의 아미노산 서열에 대해 예측을 하는 방식으로 6가지의 경우가 나오는 데 이때 그 6개의 결과에 대한 평균값을 보여주고 있다.

위의 표의 결과에서 알 수 있는 것은 훈련의 결과와 성능의 결과의 차이가 크다는 것이다. Training의 경우 Testing 결과보다 좋은 성능을 나타낸다. 하지만 Testing의 경우, 즉 트레이닝 단계에 포함되지 않은 서열에 대한 예측을 했을 경우에 그 예측 결과는 떨어지고 있다. 또한 Sov(β)의 수치가 다른 수치에 비해 상당히 높은 결과를 보이고 있다.

5. 결론 및 향후과제

이상에서 보았듯이 본 논문에서는 β -barrel 막횡단 단백질의 서열 예측을 위해 HMM을 이용해 구현한 과정과 그 결과에 대해서 보았다. 위의 결과에서 알 수 있듯이 현재 α -helix 막횡단 단백질 예측의 경우 어느 정도 결과를 보이고 있지만 β -barrel의 경우 좋은 예측률을 보여주고 있지 못한 게 사실이다. 이것은 아직 β -barrel 막횡단 단백질들의 예측에 있어서 다음과 같은 문제점들이 존재하고 있기 때문이다[14]. 첫번째는 트레이닝 셋이 β -barrel 막횡단 단백질들을 각각 대표하는 프로토타입이라고 가정하고 만들어진 트레이닝 셋의 가정이다. 따라서 그 가정이 틀린 경우 전체 시스템도 잘못된 예측 결과를 도출할 수 밖에 없을 것이다. 두 번째는, 이러한 가정이 맞았다고 하더라도 그 가정 위에서 만들어진 트레이닝 셋이 현재 12개 정도 또는 그 이하가 사용되고 있는데, 이렇게 너무 적은 트레이닝 셋에 대해 트레이닝시에 너무 많은 프리파라미터(free-parameter)들이 사용되므로 오버피팅(overfitting)을 피하기 어렵다는 점이다. 또한 α -helix 막횡단 단백질과 달리 상동성(homology)에 기반해서만 예측하는데 한계가 있다는 것이다.

결국, 문제점의 대부분은 트레이닝 셋에서부터 기인하므로 이 점은 상당히 중요한 문제가 아닐 수 없다. 따라서 트레이닝 셋에 대해 꾸준히 관심을 가지고 개선시키는 노력이 필요하다. 현재도 트레이닝 셋에 대한 문제를 해결하기 위한 노력들이 진행되고 있다. 그러므로 향후 트레이닝 셋의 개선과 상동성에 기반한 메소드에 추가적인 메소드들을 통한 개선 방법을 강구해야 할 것이다.

6. 참고 문헌

[1] Jones, D.T., Taylor, W.R., Thornton, J.M. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33: 3038-3049.

[2] Rost, B., Casadio, R., Fariselli, P., and Sander, C. 1995. Transmembrane helices predicted at 95% accuracy. Protein Sci. 4: 521-533.

[3] Rost, B., Fariselli, P., and Casadio, R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci. 5: 1704-1718.

[4] White, S.H. and Wimley, W.C. 1999. Membrane protein folding and stabiilility: Physical principles. Annu. Rev. Biophys. Biomol. Struct 28:319-365.

[5] Schulz,G.E. 2000. beta-barrel membrane proteins. Curr. Opin. Struct. Biol.,10,443-447.

[6] Cowan, S.W. and Rosenbusch, J.P. 1994. Folding pattern diversity of integral membrane proteins. Science 264: 914-916.

[7] Jacoboni,I., Martelli,P.L., Fariselli,P., De Pinto,V. and Casadio,R. 2001. Prediction of the transmemebrane regions of beta-barrel membrane proteins with a neural network-based predictor. Protein Sci.,10, 779-787

[8] Pier Luigi Martelli, Piero Fariselli, Anders Krogh, Rita Casadio: A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. ISMB 2002: 46-53

[9] Rabiner, L. R., A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77:257-286

[10] Krogh, et al., Hidden Markov Models in Computational Biology (applications to protein modeling). J. Mol. Bio. 1994. 235, 1501-1531

[11] R. Durbin, S. Eddy, A. Krogh and G. Mitchison. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.

[12] Krogh, A. 1994. Hidden Markov models for labeled sequences. In Proceedings of the 12th IAPR International Conference on Pattern Recognition, 140-144. Los Alamitos, California: IEEE Computer Society Press.

[13] Zemla, A., Venclovas, C., Fidelis, K., and Rost, B. 1999. A modified definition of Sov, a segment-based measure of protein secondary structure prediction assessment. Proteins 34: 220-223.

[14] CP Chen and B Rost 2002. State-of-the-art in membrane protein prediction. Applied Bioinformatics, Volxx, in press.