SVM과 HMM을 이용한 α-Helix 막횡단 단백질 예측

*송철환O, *유성준, **김민경, *설영주

* 세종대학교 컴퓨터공학부 ** 이화여자대학교 공학 연구소

schpeter@gce.sejong.ac.kr, sjyoo@sejong.ac.kr, minkykim@ewha.ac.kr neutrian@hotmail.com

Predicting Transmembrane α -helix protein with SVM and HMM

*Chullhwan SongO, *Seong Joon Yoo, **Minkyung Kim, *Youngjoo Seol
*School of Computer Engineering, Sejong University . **Center for Engineering Research, Ewha university

요 약

현재 바이오인포매틱스(Bioinformatics) 분야에서 가장 중요한 부분 중의 하나는 유전자 및 단백질의 구조와 기능을 정확하게 예측하는 것이다. 이는 질병 치료 및 신약개발에 유용하여 이로부터 나온 결과로부터 경제적 산업적 효과를 기대할 수 있다. 이 논문에서는 기계학습(Machine Learning)의 한 분야인 SVM(Support Vector Machine)과 HMM(Hidden Markov Model)를 결합하여 단백질의 막횡단(Transmembrane) α-Helix 단백질 지역을 예측하는 새로운 알고리즘을 개발, 구현 및 실험하였다. 그 결과 이 두 가지 알고리즘이 결합된 방식을 사용함으로써 성능을 향상 시킬 수 있음을 증명했다.

1. 서 론

본 논문은 α -helix 형태로 단백질이 막횡단에 걸치는 부분 즉, 단백질 서열(Sequence)를 예측하는 알고리즘에 대해 기술한다. 막횡단 단백질은 화학적으로 보면 소수성(Hydrophobic) 아미노산으로 구성되어 있는 특징을 갖는다. 이런 특성을 이용하여 SVM[1]과 HMM 두 모델을 가지고 막횡단 단백질을 예측한다. SVM은 패턴 인식 분야에서 좋은 결과를 보이는 기계학습 방법의 하나로 최근 바이오인포매틱스 여러 분야에도 이를 적용한 다양한 시도가 이루어지고 있다. 이 최초의 결과인 HMM만을 사용한 경우보다 SVM을 함께 사용하여 성능이 향상 되었음을 보여준다.

2. SVM 모델 기반 막횡단 지역 예측

SVM은 최근 새롭게 바이오인포매틱스의 여러 분야에 적용되고 있는 기계학습의 한 모델이다. SVM은 어떤 오브젝트즉, 아주 많이 흩어진 Vector들을 두 개의 클래스로 나누어 분류(Classification)시키는 특성을 지닌다. 이런 특성을 이용하여 20개 아미노산으로 구성된 단백질 서열을 분류하는데 적용시킨다. 이런 서열들을 Vector화하여 Kernel에 적용시키고 다양한 훈련 예로 학습을 시킨 후 새로운 단백질 서열에 대해 결과를 얻는 방법이다.

2.1 20개 아미노산 정규화

단백질 서열의 구조는 막(TM)을 중심으로 바깥지역(out)과 내부지역(in) 3개의 구조로 되어 있다. 따라서 SVM의 입력 값인 Vector는 이런 의미를 내포해야 한다. 본 논문은 이런 벡터화 과정에 4가지 모델을 사용하였다. 그 중 1개(Posterior Model)는 새롭게 모델을 세운 것이며, 3개는 이미 실험적으로 밝혀진 모델(Byod[3], Eisenberg[4], Kyte Model[5])을 적용시킨다.

새롭게 구현한 Posterior Model은 단백질의 서열 구조를 HMM모델에서 State로 간주하여 본다면 그 3개의 State를 가지고 Posterior 확률을 모든 훈련 예(Training Set)에 대하여 구한 다음에 훈련 예의 단백질 서열 중 막횡단(TM) 지역에 분포하는 20개 아미노산에 대하여 평균값을 구한 값으로 정규화 시켰다. [표 1]

[표 1] Normalized Hydrophic 20 Amino Acids of Transemembrane Protein

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	Y	W	Y
Posterior	0.86	0.82	0.65	0.67	0.87	0.86	0.65	0.80	0.51	0.87	0.83	0.73	0.78	0.66	0.44	0.64	0.65	0.66	0.82	0.80
Byod	1.37	1.12	0.17	0.16	1.93	1.05	0.74	2.20	0.19	1.70	1.39	0.43	0.51	0.35	0.30	0.33	0.89	1.81	1.58	1.01
Kyte	0.77	1.00	-1.01	-1.01	1.01	0.03	-0.91	1.97	-1.14	1.44	0.80	-1.01	-0.37	-1.01	-1.34	-0.10	-0.07	1.57	-0.14	-0.27
Eisenberg	0.62	0.29	-0.90	-0.74	1.19	0.48	-0.40	1.38	-1.50	1.06	0.64	-0.78	0.12	-0.85	-2.53	-0.18	-0.05	1.08	0.81	0.26

2.2 벡터(Vector) 화 과정

SVM이란 벡터를 어떤 기준(Hyperplane)에 의하여 분류하는 것이다. 여기서 벡터는 SVM에의 입력 값을 뜻한다. 이런 입력 값은 각각의 목적에 맞게 특성을 부여해야만 한다. 막횡단 예측 시스템 도구에서는 막횡단지역에 분포하는 아미노산들이 소수성인 특징을 사용한다. 본 논문은 다음과 같이 Training Set의 모든 단백질 서열을 벡터화 하였다.

$$f(x) = \underset{-size \le i \le size}{MAX}\left(\left(\sum_{j=i}^{i+size} x\right)/size\right) \quad \cdots\cdots\cdots\cdots (1)$$

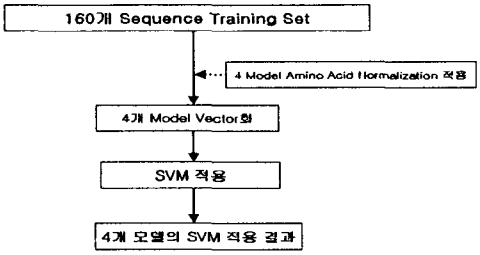
위의 식(1)에서 f(x)는 vector이고 x는 훈련 예에 있는 단백질 서열에 매칭되는 2.1에서 보여준 20개 아미노산에 대한 정규 값이다. 또한 size는 임의의 값이다. 보통 size는 막횡단에 걸치는 단백질 서열의 크기인 15~35개의 크기로 하였다. 다시 말해서 막횡단의 패턴은 15~35개 사이의 일련의 패턴으로 이루어져 나타내어진다. 본 논문은 15~35개 중 가장 성능이 좋은 15 size값으로 설정하였다. 위와 같은 과정을 통하여 모든 훈련예의 서열들을 벡터화 하였다.

2.3 SVM 적용

본 논문은 SVM 라이브러리를 제공해주는 LIBSVM을 사용하였다. 기본적으로 자바 언어로 구성된 패키지로 되어 있으며 여러 Kernel 함수를 선택하여 사용할 수 있다. 여러 Kernel을 선택하여 사용해 보았으나 Kernel 함수에 따라 그 결과가 큰 변화를 보이진 않았다. 훈련 예로 기본적으로 TMHMM에서 사용된 160개를 사용하였다. 그때의 서열 개수는 63114개이다. 또한 같은 벡터로 각각의 모델을 4번 훈련 시켰다. 또한 그 SVM의 결과는 Fitting 작업을 거쳤다. 즉, 35개 이상의 막횡단 단백질 서열 지역이 over-fitting된 경우이므로 두개로 나누거나 35개 이하로 잘라주는 역할을 한다.

2.4 SVM 모델 구조 및 평가

[그림1]은 이제까지 본 논문에서 시도된 SVM모델에 대해 설명한 것이다. 우선 160개의 막횡단 단백질의 서열을 포함하는 훈련 예를 가지고 4가지 모델 각각으로 벡터화



[그림 1] SVM 모델 구조

하는 [표1]을 적용하는 과정을 거친다. 그 후 SVM에 적용시켜서 결과를 생성한다.
기본적으로 SVM은 이진법적 분류 특성을 갖는 바 현재 구현된 예측 시스템은 topology예측(in/out)은 수행하지 않고 막횡단 지역인지 아닌지 만을 수행한다. 따라서 TMHMM에서 성능 평가 방법 중 topology 예측을 제외한 3개의 평가 기준에 따라 성능을 시험하였다.

[표2] SVM 적용 결과

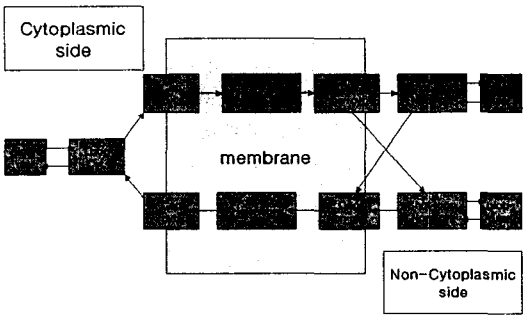
	Posterior	Byod	Kyte	Eisenberg
Correct location (%)	48	44	47	55
Single TM sensitivity(%)	87	90	92	92
Single TM specificity(%)	94	90	84	94

Correct location : 방향성에 상관없이 TM 부위를 정확히 예측 비율
Single TM Sensitivity :예측된 TM/true TM
Single TM specificity : 실제 정확히 예측 된 TM/모든 예측된 TM

[표2]에서는 각각의 4개의 모델에서 얻어진 결과를 보여준다. 그 결과들은 거의 비슷한 성능을 나타낸다. 각각의 평가 방법에 따라 차이가 나지만 [표1]의 정규화는 비슷한 특징을 가지고 있다는 것을 알 수 있다. 또한 각각의 모델에서는 서로 예측하지 못한 막횡단 단백질 서열들을 서로 보완해주는 역할을 하고 있다.

3. HMM 모델을 적용하여 막횡단 지역 예측

HMM은 TMHMM[1]에서 보여준 모델을 적용시켰다. 이 논문에서는 크게 7개의 모델을 가지고 헬릭스 코어부분, 헬릭스 코어부분의 양끝부분, 세포막 안쪽의 루프, 세포막 밖의 두개의 루프로 구성된 7개의 유형의 state를 가지고 있으며 서로 순환 한다[그림2].



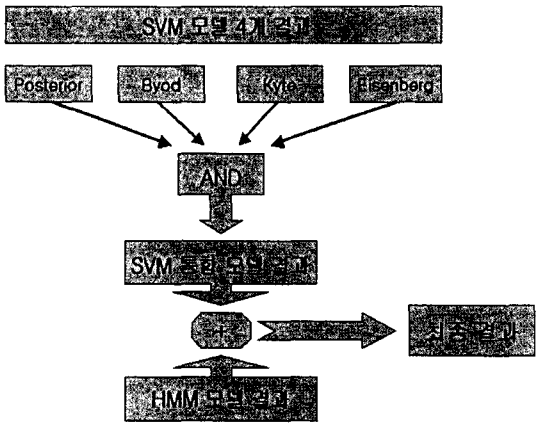
[그림2] α -helix의 HMM 모델 구조

[그림2]에서는 모델의 전체 구조를 보여준다. 모두 7개의 state를 보여주고 있으며 또한 아미노산 분산확률과 각각의 state의 트렌지션 확률도 같다.
HMM에서 훈련은 기본적으로 HMM에서 모델 훈련을 하

는데 Maximum likelihood estimation을 하는 표준 방법인 Baum-Welch re-estimation을 사용한다[6]. 여기서는 각 서열에 topology에 해당하는 TM, in, out으로 레이블링된 서열로 훈련을 한다. 훈련이 끝난 다음에는 가장 큰 확률을 지닌 패스를 찾기 위해 Viterbi algorithm을 사용한다.

4. SVM+HMM 알고리즘을 이용한 예측

2.3장에서 SVM 4개의 모델을 가지고 그 결과를 보았다. 통합함으로써 예측 성능의 신뢰성을 높일 수 있기 때문이다. 또한 최종적으로 통합된 SVM 모델과 HMM 모델은 상호 보완적으로 그 성능을 향상 시킨다[그림 3].



[그림 3] SVM+HMM 통합 구조

5. 평가

2.4에서 보여주고 있는 평가방법을 그대로 사용한다. 그리고 HMM모델의 결과를 보여주고 또한 SVM의 4가지 모델을 합한 결과에서 HMM과 모델을 합한 결과를 가지고 평가한다[표3].

[표3] HMM과 SVM+HMM의 결과 비교

예측 알고리즘	Correct location (%)	Single TM sensitivity(%)	Single TM specificity(%)
HMM	67	95	96
SVM+HMM	75	97	96

HMM과 SVM에서 비교해보면 평균적으로 HMM에서 성능이 높은 것을 알 수 있었다. 하지만 SVM과 HMM의 두 개의 모델을 가지고 합한 결과를 보면 Correct location에서 성능이 HMM모델을 단독으로 사용하였을 때 보다 약 8% 향상 되었음을 알 수 있다. 그 이유는 topology의 정확성이 훨씬 더 높아졌음을 표[3]에서 볼 수 있다.

6. 결론

이제 까지 본 논문은 단백질에서 α -helix 막횡단 지역의 서열들을 예측하는데 SVM과 HMM를 이용한 예측 도구 개발 결과를 보았다. SVM과 HMM은 기본적으로 비슷한 결과를 보인다. 하지만 두개의 모델을 합쳤을 경우에는 하나의 방법을 사용한 결과보다 더 향상시키는 결과를 얻을 수 있었다. 상호 보완적 역할을 하기 때문일 것으로 생각된다. 다시 말해서 두개 모델이 예측한 막횡단 단백질 서열 지역이 서로 검증할 수 있고 또 서로 찾지 못한 지역을 보완하여 그 막횡단 단백질을 찾아서 더욱더 성능을 향상 시키기 때문이다. 앞으로는 기능상으로는 topology를 예측하는 기능과 α -helix가 아닌 β-barrel의 형태로 막횡단 하는 단백질을 예측하는 기능을 추가해야 할 것이고 성능면에서는 over-fitting 문제를 해결하고 의 형태로 아닌 예측하지 못하는 부위를 찾아서 더욱더 성능을 향상 시켜야 할 것이다.

7. 참고 문헌

[1] Joachims, T. (1999) Making large-Scale SVM Learning Practical. In: Sch_lkopf B, Burges C, Smola A, editor. Advances in Kernel Methods-Support Vector Learning MIT Press, p 41-54

[2] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.*Journal of Molecular Biology*, 305(3):567-580, January 2001.

[3] Boyd, D., Schierle, C. and Beckwith, J. (1998) How many membrane proteins are there? *Protein Sci.*, **7**, 201-205.

[4] Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125-142.

[5] Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105-132.

[6]Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77(2):257□ 286.